COMPARISON OF ORTHOLOGS ACROSS MULTIPLE SPECIES BY VARIOUS STRATEGIES

BY

HUI LIU

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Biophysics and Computational Biology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Doctoral Committee:

Professor Eric Jakobsson, Chair, Director of Research
Professor Gene E. Robinson
Associate Professor Saurabh Sinha
Assistant Professor Jian Ma

Abstract

Thanks to the improvement of genome sequencing technology, abundant multi-species genomic data now became available and comparative genomics continues to be a fast prospering filed of biological research. Through the comparison of genomes of different organisms, we can understand what, at the molecular level, distinguishes different life forms from each other. It shed light on revealing the evolution of biology. And it also helps to refine the annotations and functions of individual genomes. For example, through comparisons across mammalian genomes, we can give an estimate of the conserved set of genes across mammals and correspondingly, find the species-specific sets of genes or functions.

However, comparative genomics can be feasible only if a meaningful classification of genes exists. A natural way to do so is to delineate sets of orthologous genes. However, debates exist about the appropriate way to define orthologs. It is originally defined as genes in different species which derive from speciation events. But such definition is not sufficient to derive orthologous genes due to the complexity of evolutionary events such as gene duplication and gene loss. While it is possible to correctly figure out all the evolutionary events with the true phylogenetic tree, the true phylogenetic tree itself is impractical to be inferred. Furthermore, evolutionary orthology does not necessarily have a strict correspondence to function similarity.

Kruppel type zinc finger genes are important transcription factor families in eukaryotic species. Important as it is, the evolution of this zinc finger family is not completely clear yet. For example, the vertebrate roots of the KRAB-ZNF family in particular and of polydactyl ZNF genes in general, remain somewhat mysterious. In addition, due to its repeating gene structure and the fact that they often reside as clusters, these genes can be difficult to model correctly with common gene-finding tools. Furthermore, due to the repeating gene structures, conventional tools like BLAST lack the necessary sensitivity to successfully define orthologous relationships.

This research presents a novel and species-specific way of defining orthology for zinc finger genes-‘fingerprint’ alignment, hundreds of lineage-specific genes in each species and also hundreds of orthologous groups are found. Most groups of orthologs displayed some degree of fingerprint divergence between species. Focusing on the dynamic KRAB-ZNF subfamily, only three genes conserved between mammals and nonmammalian groups are found. These three genes, members of an ancient familial cluster, encode an unusual KRAB domain that functions as a transcriptional activator. Evolutionary analysis confirms the ancient provenance of this activating

KRAB and reveals the independent expansion of KRAB-ZNFs in every vertebrate lineage. Most human ZNF genes, from the most deeply conserved to the primate-specific genes, are highly expressed in immune and reproductive tissues, indicating that they have been enlisted to regulate evolutionarily divergent biological traits.

Notably, the honeybee has been successfully used as a model to study social-behavior. The honeybee has highly-socialized colony and exhibits varieties of social behaviors such as foraging. Simultaneously, it is relatively simple and easy for honeybee to be manipulated comparing to other social animals. However, it is still unclear about the evolutionary relationship between honeybee and other social animals like human. Specifically, is there a conserved common genetic basis for social behavior between honey bee and human?

Based on conventional ortholog data defined by whole-sequence comparison, ortholog distribution patterns are compared for sets of aggression-related honey bee genes. We found that for one particular stimulus, response to alarm pheromone, the set of honey bee genes differentially expressed in the brain contains disproportionately large numbers of genes also found in mammals, including humans. Functionally, a large number of the human counterparts of these genes are important for regulating protein folding, a process whose misregulation is prominently implicated in human neurodegenerative disease. Moreover, the human counterparts are also predicted to be co-regulated similarly to the bee genes that respond to alarm pheromone, even though alarm pheromone is a highly species-specific signal. These results suggest surprisingly strong similarities in socially responsive genetic circuits common to honey bees and mammals.

Acknowledgments

Grateful thanks to my advisor: Eric Jakobsson, Professor Emeritus of Biochemistry, Molecular and Integrative Physiology, Biophysics and Computational Biology, Neuroscience Program

Grateful thanks to (for support and guidance to complete the works done in this thesis): Lisa Stubbs, Professor of Cell and Developmental Biology. Gene E. Robinson, Professor of Entomology and Neuroscience Program

Grateful thanks to (for guidance and help as my defense committee members): Jian Ma, Assistant Professor of Bioengineering, Biophysics and Computational Biology, and Computer Science, Saurabh Sinha, Associate Professor of Computer Science, Entomology and Biophysics and Computational Biology

Thanks for useful discussion with members of the Institute for Genomic Biology themed “Gene Networks in Neural & Developmental Plasticity”. Thank Elbert Branscomb for many useful discussions related to the work of Chapter 2, and Yi Xing for providing the processed RNA-seq data set. Thank Stuart Huntley for carrying out the early pilot work that inspired the work of Chapter 2. The works of Chapter 3 were partially supported by an NIH Director’s Pioneer Award 1DP1OD006416 to Gene E. Robinson and National Science Foundation Award 0835718 “Hierarchical Modularity in Evolution and Function” to Eric Jakobsson.

Table of Contents

Chapter 1. General Introduction **1**

Chapter 2. Explore the Evolution of C2H2 Zinc Finger Gene by Gene Model Construction and Fingerprint Alignment Orthology **6**

Chapter 3. Conservation Across The Metazoa Of Differentially Expressed Genes Associated With Honeybee Behavioral Phenotypes **25**

Chapter 4. Future Perspectives **40**

Bibliography **42**

Tables and Figures **58**

APPENDIX A – Supplementary Tables **99**

APPENDIX B – Supplementary Data Files **102**

Chapter 1. General Introduction

With the advancement of genome sequencing technologies and analyzing methods, novel genomes are sequenced at an increasingly fast pace. According to NCBI, there were 56 new eukaryotic genomes sequenced in 2013 alone, and the total number of available eukaryotic genomes is more than 100 by now. There are thousands more sequenced genomes to be counted if one includes prokaryotic genomes[1]. Consequently, comparative genomics quickly became an important field of biological research, especially because it is revolutionizing our ability to explore biological evolution. In addition, it helps us to acquire a greater knowledge about the cross-species difference and commonalities[2].

Orthology

Orthology can be thought of as having three different dimensions: 1) homology, similarity of sequence; 2) phylogeny, descent from a common ancestor without intervening duplication; 3) ontology, commonality of function. Various ortholog identification methods that consider different dimensions of these are developed and these are briefly reviewed in the following three sections.

Sequence-based Methods

COG[3] is the most classical method based on BLAST symmetric best hits(SymBets)[4]. Based on pairwise BLAST comparison of all sequences, this method tries to identify triangles of mutually consistent, genome-specific best hits and then merge those triangles with a common side to form ‘Clusters of Orthologous Groups’ (COGs). The groups produced include orthologs from different lineages and, in many cases, paralogs from the same lineage. There were improvements with regard to the performance of the algorithm[5] not long ago but it remains focused primarily on prokaryotes and lacks a lot of eukaryotic species[6]. InParanoid, which is also based on the ‘SymBet’ idea, presents the new concepts of in- and out-paralogs[7]. For pairwise genomes, this method first identifies orthologous pairs, which serve as seeds in the orthologous groups. Then in-paralogs are explicitly detected and added based on the assumption that sequences from the same species that are more similar to the seed ortholog than

to any sequence from other species are in-paralogs belonging to the same group of orthologs. In contrast to COG, InParanoid emphasizes eukaryotic species and has been continually updated[8,9]. The most updated version InParanoid8 now contains 273 organisms, which makes it a comprehensive eukaryotic ortholog resource. Since InParanoid only presents results of pairwise genome alignments, MultiParanoid[10] has been developed. In addition to InParanoid, several other 'SymBets' methods are also formulated[11,12]. Although orthology based on sequence information can be inconsistent from the evolutionary true orthology, 'SymBets' approach is among the most popular approaches on account of its computational advantage and generally good accuracy.

Phylogenetic Tree Based Methods

As the definition of 'ortholog' is based on evolutionary events, methods based on phylogenetic trees have been the most classic method of ortholog identification: a phylogenetic tree of homologous genes are built and are compared to species tree, so that evolutionary events, like gene duplications, can be discovered, and correct homology type can be inferred[13]. Such a procedure or method is called "tree reconcilation"[14]. A couple of methods and databases have been established based on this method [15–17]. Interestingly, one of the methods, RIO, calculates orthologs[16] using multiple alignment profiles(based on the Pfam database), instead of the conventional single-sequence-based orthologs. However, the phylogenetic tree method is not widely used due to the following reasons: first, the precision and reliability of species tree and phylogenetic tree of homologs is questionable in many cases; second, the computation cost is expensive, especially for genome-wide analysis and applications; third, for prokaryotes, the horizontal gene transfer events makes the reconciliation based on single species tree difficult. Moreover, the basis of the phylogenetic tree inference — the multiple sequence alignment — is impractical to be optimally solved due to exponential time complexity and the variability of scoring matrices used [18–20]. Efforts have been made, however, to alleviate these problems: for example, to avoid the explicit usage of a species tree[21], or try to find more efficient implementations[22]. Probabilistic approaches are also used for the improvement of tree-based methods[23,24].

In addition, there are hybrid approaches combining 'SymBets' and tree methods. For examples, HomoloGene(http://www.ncbi.nlm.nih.gov), TreeFam[25], OrthoDB[26] and Ensembl Compara[27].

Ontology-Functional Aspects of Orthology

In addition to sequence similarity, the functional similarity of orthologous genes is an important concern. It is sometimes assumed that evolutionarily orthologous genes should have the same function[4][3]. This is often true. However, if there are a lot of duplications or gene loss or if the evolutionary distance is large, the functions might not be conserved any more[28]. Non-orthologous genes can employ the same function[29] and orthologous genes' functions can be diverged[30]. Gene ontology is a widely used tool for describing 'orthologous' biological functions that are shared across eukaryotic genes [31]. And issues with orthology have a deep connection to issues in gene ontology. An recent study based on the functional similarity using Gene Ontology indicates that the functional similarity of paralogs within species might be higher than orthologs. Notably, it is the cellular context in addition to the sequence that affects the function evolution[32].

Domain/Motifs-Another Dimension of Orthology

"Domains" (or "motifs", these two terms will be used interchangeably in this dissertation) are subunits of proteins that are relatively independent both structurally[33] and evolutionarily[34]. Various protein motif databases and analysis tools — Pfam[25], SMART[35], PROSITE[36], Gene3D[37], TMHMM[38], SUPERFAMILY[39] — are developed and continually updated. And InterPro[40] is a comprehensive interface that provides access most of these protein motif databases. It is very common to make comparisons based on similarity (homology) of complete genes or protein gene products. It is also common to identify particular bases or residues that are of critical importance to molecular function, based on changes in molecular function in response to particular mutations[41–43]. However, the concepts of orthology and paralogy can be extended to motifs or domains as well as to entire genes and proteins[4]. This extension is necessary because genes do not evolve solely by mutations, insertions, and deletions in single sites. Often entire sections of a gene will appear or vanish within a lineage of descent. Indeed, gene fusion/fissions are important evolutionary events that affect orthology[4]. Further, genes are often characterized by motifs. For example the voltage-gated ion channel family exhibits enormous diversity with respect to functional roles and modes of regulation[44].

The degree of mixing and matching of domains in genes and their protein gene products, and the organization of those domain-containing proteins in searchable databases, has resulted in a situation where the linear format of

the traditional literature review article is inadequate to describe family relationships of proteins. For example, using a sequence segment from a randomly selected potassium ion channel as input to InterProScan, more than one hundred distinct domain organizations are found. Some of the domain identifications are particular to potassium ion channels but most are not, including, for example, cyclic nucleotide-binding domain (IPR000595), BTB/POZ fold (IPR011333), extracellular solute-binding protein, family 3 (IPR001638), NAD(P)-binding domain (IPR016040), and zinc finger, CCHC-type (IPR001878). These domains represent the various ways in which potassium ion channels and many other proteins can be regulated. Indeed, hERG[45], which plays a critical role in heart rhythm, was found to be composed of both potassium ion channel and cyclic-nucleotide-binding domains. In addition, domain-composition differences also exist for a lot of enzyme families: e.g., the DnaG-like primases of bacteria and archaea[4], the thiamin pyrophosphate (TPP)-dependent enzymes and glycosyl hydrolases with TIM barrel fold[46]. Alternative splicing is also an important way to create protein diversity[47] such as creating variants of MHC molecules[48]. Domain duplication is also discovered in some cases[46]. A prime example of a ubiquitous domain that is inserted into proteins in a great variety of ways, which is chosen to be studied in this dissertation, is the zinc finger motif. Thus, although protein families are often designated by either an active domain, such as potassium ion permeation or potassium ion transport, or by a regulatory domain, such as cyclic nucleotide binding, neither designation is definitive or adequate, because different families defined in this manner are not isolated. Rather each "family" needs to be seen as a common characteristic, generally represented in the sequence by a functional domain, or a combination of functional domains, common to the family members, that has the ability to become attached in the course of molecular evolution to any of a number of other functional domains. This process of domain mixing has been referred to as the "horizontal dimension" of protein evolution[49].

To summarize about orthology, there are various levels to define orthology and the most appropriate level of ortholog analysis depends on the system under study and the questions asked. In the next two chapters of this dissertation I will present two detailed studies.

The first study focuses on orthology at the domain level (zinc fingers). Conventional sequence-based approaches such as BLAST on complete genes and gene products fail when applied to zinc-finger gene orthology, ontology, due to highly similar and repeated gene substructures within zinc finger gene family. Thus a domain-based approach must be used, which will be explored in this dissertation.

The second study focuses on orthology at the gene product (protein level) to understand cross-species comparisons of the genomic basis of social behavior. Based on conventional ortholog data defined by whole-sequence comparison[9], ortholog distribution patterns are compared for sets of aggression-related honey bee genes[50]. Specifically, we analyzed the extent to which the differentially expressed genes are conserved across the Metazoa. We also employed an alignment-free similarity measure, the D2z method[51], to analyze the promoter regions of human orthologs to infer the likelihood of human patterns of co-expression similar to those observed in the honey bee experiments.

Chapter 2. Explore the Evolution of C2H2 Zinc Finger Gene by Gene Model Construction and Fingerprint Alignment Orthology[1]

Abstract

While many vertebrate transcription factor (TF) families are conserved, the C2H2 zinc finger (ZNF) family stands out as a notable exception. In particular, novel ZNF gene types have arisen, duplicated, and diverged independently throughout evolution to yield many lineage-specific TF genes. This evolutionary dynamic not only raises many intriguing questions but also severely complicates identification of those ZNF genes that remain functionally conserved. To address this problem, we searched for vertebrate "DNA binding orthologs" by mining ZNF loci from eight sequenced genomes and then aligning the patterns of DNA-binding amino acids, or "fingerprints," extracted from the encoded ZNF motifs. Using this approach, we found hundreds of lineage-specific genes in each species and also hundreds of orthologous groups. Most groups of orthologs displayed some degree of fingerprint divergence between species, but 174 groups showed fingerprint patterns that have been very rigidly conserved. Focusing on the dynamic KRAB-ZNF subfamily—including nearly 400 human genes thought to possess potent KRAB-mediated epigenetic silencing activities—we found only three genes conserved between mammals and non-mammalian groups. These three genes, members of an ancient familial cluster, encode an unusual KRAB domain that functions as a transcriptional activator. Evolutionary analysis confirms the ancient provenance of this activating KRAB and reveals the independent expansion of KRAB-ZNFs in every vertebrate lineage. Most human

[1] This chapter appeared in its entirety in the Journal of Genome Biology and Evolution is referred to later in this dissertation as "Liu H, Chang L-H, Sun Y, Lu X, Stubbs L (2014) Deep vertebrate roots for Mammalian zinc finger transcription factor subfamilies. Genome Biol Evol 6: 510–525." And it is Available as: http://gbe.oxfordjournals.org/content/6/3/510.short.

All computational analyses and part of the paper writing and discussions are done by Hui Liu. Main project idea, discussions and part of the analysis is credited to Professor Lisa Stubbs. RNA-seq, quantitative PCR, In Situ hybridization, immunohistochemistry and related primary analysis of experimental data are credited to Li-Hsin Chang, Younguk Sun and Xiaochen Lu

ZNF genes, from the most deeply conserved to the primate-specific genes, are highly expressed in immune and reproductive tissues, indicating that they have been enlisted to regulate evolutionarily divergent biological traits.

Introduction

Most eukaryotic transcription factors (TFs) are members of ancient protein families, and many are conserved across divergent evolutionary lineages. However, this latter generalization does not hold universally true, and the C2H2 zinc finger (ZNF) family stands out as a particularly significant exception. At several points in evolutionary history, novel gene types have arisen to encode proteins in which DNA-binding ZNF motifs are tethered to different types of chromatin-interacting or "effector" domains. Some of these innovations have subsequently been expanded by duplication into large cohorts of lineage-specific genes[52,53].

ZNF proteins that function as TFs typically contain an array of two or more tandemly arranged C2H2 motifs; each ZNF in such polydactyl fingered or "polydactyl" proteins can bind three adjacent nucleotides at target sites with amino acids in positions −1, 2, 3, and 6 in the alpha helical region of each motif playing the most critical DNA-recognition roles[54,55]. Adjacent motifs influence each other's DNA binding, creating a complex "code" that links the pattern of DNA-binding amino acids in a protein to target-site preferences in DNA[56,57]. In the following discussion, we will refer to the pattern of DNA-binding amino acids within a polydactyl ZNF array as a protein's "fingerprint." It stands to reason that ZNF proteins with similar fingerprints should recognize similar DNA sequences, while even closely related proteins with divergent fingerprints should preferentially interact with different recognition sites in DNA. An extreme example of this type of fingerprint divergence is provided by PRDM9, an ancient protein that binds hotspots of meiotic recombination. PRDM9 orthologs encode proteins that are similar in overall sequence, but that nevertheless define hotspots uniquely in every species using ZNF arrays that have been positively selected for fingerprint divergence[58–62].

Interestingly, although PRDM9 is unique in invertebrate genomes, this single gene's descendants have expanded to form the largest ZNF subfamily in mammalian genomes[63]. The human genome encodes hundreds of these so-called KRAB-ZNF genes, encoding proteins in which arrays of tandem ZNF motifs are tethered to an N-terminal effector domain called the Krüppel-associated box or KRAB[64–66]. The canonically structured mammalian "KRAB A" domain interacts with a universal cofactor, KAP1, which recruits histone deacetylase

complexes to the ZNF-binding sites, and KRAB-ZNF proteins are thus thought to act as potent transcriptional repressors[67–70]. However, the vertebrate roots of the KRAB-ZNF family in particular, and of polydactyl ZNF genes in general, remain somewhat mysterious. For example, it is not known which human polydactyl proteins are conserved in structure and function in other vertebrate species or which among the otherwise conserved proteins, like PRDM9, might have been selected especially for DNA-binding diversity.

To address these questions, we used methods that we applied previously to identify mouse, dog, and primate genes[66,71] to collect consistent sets of polydactyl ZNF gene models from the opossum, chicken, zebra finch, lizard, frog, and updated mouse genomes. From these models, we extracted and aligned fingerprint patterns to identify proteins with similar or divergent DNA binding capacities. We identified hundreds of polydactyl ZNF loci in every genome including more than 100 predicted novel mouse genes, but surprisingly few encoding proteins with fingerprint patterns that are conserved between eutherians and other evolutionary groups. Notably, the subset that is deeply conserved includes only three KRAB-ZNF genes, all of which map to a single familial cluster. These ancient genes share certain features that are unusual in mammalian genomes, including a noncanonical KRAB domain sequence that does not bind KAP1 and functions as a transcriptional activator [72,73]. These and other findings suggest a history in which the KRAB-ZNF proteins expanded and diverged independently in every vertebrate lineage including amphibians, where they expanded without KAP1-interacting capabilities, very possibly as activating TFs.

The rigid preservation of DNA-binding domains suggests that the conserved polydactyl ZNF genes have been stably integrated into essential regulatory relationships. Strikingly, however, the most deeply conserved genes are expressed at highest levels in human tissues that are the least conserved in structure and function, including placenta. Our results identify hundreds of novel polydactyl ZNF genes of both deeply conserved and lineage-specific types, providing new clues to the history and root functions of this dynamic TF family.

Methods

Genome Searches and Initial Data Analysis

Human KRAB-A, KRAB-B, KRAB-b, KRAB-C, KRAB-L, BTB/POZ, SCAN and FINGER HMM matrices

are retrieved from previous analysis[66]. Chicken KRAB-A-containing protein sequences from NCBI (sequences are trimmed according to HMMER result to get KRAB-A only sequences) and Pfam KRAB, SCAN and BTB/POZ sequences are also retrieved. Sequence alignments for each motif-type were generated by using CLUSTALW 2.0.10[74] and submitted to the HMMER(hmmer.org) profile HMM matrix building tool 'hmmbuild' to generate matrices(and processed by 'hmmcalibrate'). These matrices were used by the HMMER search program to identify all putative motif matches in a full six-frame translation of all the chromosome sequences of frog, lizard, zebra finch, chicken, opossum and mouse (xenTro3, anoCar2, taeGut1, galGal3, monDom5, mm9 from UCSC genome browser[75]). EST data are also retrieved from UCSC genome browser. They are grouped if overlapped. And then used the grouped EST data are mapped to motif hits as annotations (Appendix B Datafile B1,B3). An e-value cutoff of 0.001 is used. HMMER hits with stop codon in frame are also filtered out. For overlapped hits, the hit with lower e-value is kept.

Gene Model Construction

Gene model structures were constructed by the following procedure:

1. Grouping motifs with no genome gap between (bridged gap is ignored since the order and orientation of either side of the gap is known). Then motifs with distance larger than 30Kb were further separated (if two motifs are zinc fingers, this threshold is stringent to 1Kb). Separating cofactor motifs if they are at the 3' end of zinc fingers.
2. For each cluster, considering all possible combinations of the upstream cofactors, especially the 6 main subfamilies of KRAB ZNF families[76], generate all possible transcripts.
3. Extending exon boundaries maintaining canonical intron splice sites (GT-AG, AT-AG, and GC-AG) and find the nearest start, stop codons. And make sure no stop codon in frame.
4. (For mouse only), models are checked comparing to existing Ensembl gene models and refined. Fragmented models are glued together if they are in the same Ensembl gene model.

Finally, gene models that have at least 2 zinc finger motifs are kept. For each gene, only the longest transcript is kept.

Zinc Finger "Fingerprint" Extraction

For mouse, chicken, opossum, zebra finch, lizard and frog, zinc finger motif sequences are retrieved based on HMMER search results. Then they are aligned with a standard finger sequence ("YE**C**SE**C**GKSFSRSSHLIV**H**QRI**H**TGERP", a zebra finch C2H2 zinc finger HMMER hit with e-value 5.8e-21). Amino acid immediately precede alpha-helix and the 2^{nd},3^{rd} , 6^{th} amino acid (right before Histidine) residues are retrieved as the "Fingerprint" (e.g, the "Fingerprint" is "RSHV" for the standard finger above)[77]. Previously established human and Dog ZNF gene models[66][71] were also used to extract Fingerprint data for cross-species comparison.

Fingerprint Alignment and Clustering of Zinc Finger Genes

Pairwise alignment of the 4-aa fingerprint sequences from genes from the 8 species (frog, lizard, zebra finch, chicken, opossum, dog, mouse, and human) was carried out using the Global Alignment Algorithm with gap penalty=1, mismatch penalty=1, match penalty=-2, similar penalty = -0.5 (2 out of 3, or 3 out of 4 positions in the Fingerprints are the same), closematch penalty= -1 (only the 2^{nd} residue is different). The scores were first normalized by sequence length, then were scaled as $Score(x,y) = Score(x,y) - Score(x,x)$ (so that the score is always non-negative and equal to zero if and only if two fingerprint sequences are identical). The normalized scores were used as a distance matrix and served as input for an agglomerative hierarchical clustering. The clustering was done in R using average linkage criteria (http://stat.ethz.ch/R-manual/R-patched/library/stats/html/hclust.html). The genes were grouped by cutting the clustering tree at a height of 2 (height of 1~5 are tried and 2 is chosen by considering both the discrimination power and stringency). Then, for each group, a multiple alignment using UPGMA[78] guide tree was generated. After an initial alignment, we identified many conserved groups without a human ortholog included, even though human orthologs are well known to exist. These human genes, which were missed in our previous study due to the stringent requirement for at least 3 tandem ZNF-encoding motifs in the sequence in those genome scans, were retrieved from NCBI for a second round of fingerprint alignments. We included all ZNF-containing isoforms recorded for those proteins, so that they can be seen in multiple fingerprint alignments. But for final counts of gene and ortholog numbers we included only one isoform, encoding the longest protein, for each gene.

Finding Orthologs by Reciprocal Best Hit using BLAST and Fingerprint Alignment

Gene ‘a’ in species A and Gene “b’ in species B are defined as reciprocal best hit here if Gene ‘b’ has the highest score with respect to Gene ‘a’ in species B, and it is also true vice versa. For comparison reason, both Fingerprint alignment and BLAST are used for scoring here. For fingerprint alignment, the normalized pairwise alignment score is used (same used for generating the distance matrix, see previous section). For BLAST, e-value is used for ranking the hits. As there are too many top hits with the same e-values in some cases, at most 10 top hits are retained for each query when doing BLAST. The e-value and coverage of reciprocal best hit are the averages of the corresponding values of two one-way BLASTs.

Tree Construction and Display

The tree of KRAB A motifs was generated using PhyML by the NNI search method, with SH-like branch support[79] which is of the range 0~1.0: the larger score is more significant. Tree Graphs were generated using Python ete2 package[80]. To obtain information regarding the history of all gene-linked human KRAB A domains, we used all human KRAB A sequences identified in previous studies regardless of C2H2 motif association[66]. For all other species we used only KRAB A domains included in ZNF gene models that are described here.

Generation of Consensus Sequence of KRAB A domain

KRAB-A sequence alignment of Human, Opossum, Chicken and Frog from tree construction are used. The sequence logos of Figure 3 are generated using WebLogo[81].

RNA-seq and Cluster Analysis

RNA-seq expression data, including data from the public BodyMap project (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513/, last accessed February 27, 2014) and supplemented with published expression data from amnion, chorion, and decidua of human term placenta, were kindly provided in the form of processed uniquely mapped log2 fragment per million reads (FPKM) values by Dr Yi Xing (University of California, Los Angeles). The processing steps and sources of raw data are cited in the Xing laboratory’s recent article[82]. We removed genes with FPKM values that were not at least 1 in any tissue and used Cluster 3.0

Software[83] to generate the heat maps from data centered to the median of each gene's expression levels for Hierarchical clustering with Average Linkage.

RNA Preparation and Quantitative PCR

Animal work described in this study was carried out in strict accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health. The protocol was approved by the Institutional Animal Care and Use Committee of the University of Illinois (Animal Assurance Number: A3118-01; approved IACUC protocol number 11030). RNA was prepared from snap-frozen dissected mouse embryos collected from timed matings at various stages of development and purified using Trizol (Invitrogen) extraction. cDNA was generated using Superscript III Reverse Transcriptase (Invitrogen, CA) according to manufacturer's instructions. Interexonic qRT-PCR primers for Zfp282 (forward: 50-TGACTGCAGACACAGGAACAG-30, reverse: 50-CTCTGCCAAATCCTGCTGGT-30) and Zfp777 (forward: 50-TTCCCAAGGTTCCTGTCACATTC-30, reverse: 50-CGTCTCACCCTCCTCAGAATC-30) were synthesized from IDT(Coralville, Iowa). Reaction was carried out using Power SYBR Green PCR master mix (Applied Biosystems, Foster City, CA) on the ABI7900HT system. Expression levels were calculated relative to the average expression of two housekeeping genes, Succinate dehydrogenase complex, subunit A (Sdha: accession number BC011301) and Tyrosine 3-monooxygenase/tryptophan 50monooxygenase activation protein, zeta polypeptide (Ywhaz, NM011740), as described[84,85].

In Situ Hybridization and Immunohistochemistry

Mouse embryos, placenta, and yolk sac were isolated at embryonic day 12.5 (E12.5) or E16.5 after timed matings. Chicken embryos were collected after incubation of freshly fertilized eggs at 37 ℃. Dissected embryos were fixed in fresh 4%paraformaldehyde (PFA) and embedded in paraffin. Human normal term placenta was obtained from an anonymous donor at the Carle Foundation Hospital (Urbana, IL) and provided as PFA-fixed, dissected maternal and fetal tissue segments that were subsequently embedded in paraffin. Tissues were cut into 5-mm sections using a Leica RM2155 microtome and Super Plus charged slides.

For mouse in situ hybridization (ISH), probe sequences were generated from sequenced cDNA clones from the

IMAGE consortium: mouse LIBEST_005352 clone 9530039E11 (accession number BY722098) and chicken clone LIBEST_011205 CSEQRBN13ChEST197h24 (accession number BU448580). Mouse primers for PCR were as follows: forward, 50-AGGACAGACCAGAATGCATC-30 and reverse, 50-CGAAGCTACTGACAAGGTGT-30; the chicken probe was generated using primers: forward, 50-ACAAGACAACGCACAATGCC-30 and reverse, 50-TATCTGGAAGACCGTGTTGC-30. Probe sequences were labeled and hybridized essentially as described[86]. Immunohistochemistry (IHC) was also carried out essentially as described [86]. We used primary antibodies: ZNF282 (AVIVA, Rb38361) and ZNF777 (AVIVA, Rb32569) diluted 1:200(5 µg/ml). The sections were incubated with primary antibodies overnight at 4 ℃, then washed and incubated with secondary antibody Alexa Fluor 594 goat anti-rabbit IgG (Invitrogen), 1:200 diluted (1 µg/ml). The results were imaged with a Nanozoomer Scanner (Hamamatsu), Zeiss Apotome Fluorescence microscope, and Zeiss Confocal Microscope LSM 710.

Result

Identification of Polydactyl ZNF Genes in Sequenced Genomes

We used methods based on those described previously [66,71] to identify potential. ZNF coding genes in the M. musculus (mouse, mm9 genome build), Mo. Domestica (opossum, monDom5), G. gallus (chicken, galGal3), T. guttata (zebra finch, taeGut1), A. carolinensis (lizard, AnoCar2.0), and X. tropicalis (frog, xenTro3) genomes. Of these assemblies, only the mouse genome is finished sequence. With the expectation that many genes could be fragmented in the unfinished genomes, we built gene models requiring only two closely spaced ZNF HMMER matches, rather than three tandem ZNFs as we had in the previous human, dog, and primate genome analysis. We also scanned each genome for HMMER models corresponding to the BTB/POZ, SCAN, and KRAB effector domains and included exons encoding those domains into ZNF gene models where possible as previously described[66,71].

We gathered substantial numbers of polydactyl ZNF encoding ORFs from every species including members of all subfamilies defined by association with the common known effector domains (Table 1). These gene model sets are very likely to include recent pseudogenes; we examined overlap with the other annotated gene sets for additional model support.For the 1,194 mouse polydactyl ZNF models, we identified 799 overlapping with known genes

and/or ENSEMBL gene models; the counts of previously annotated mouse and human gene 70 are roughly comparable (Table 1).

However, we also found 210 mouse loci with ORFs encoding five or more contiguous ZNF motifs but without known gene assignment or ENSEMBL models; 9 predicted genes share fingerprints with annotated human genes and models in other species and are likely orthologs (Table 2). Furthermore, 131 of these novel mouse models overlap mouse EST sequences, most of which are derived from oocytes, preimplantation embryos, or dissected tissues from midgestation embryonic stages (Appendix A Table A1; examples in Table 2, 3). Many of the unknown genes are found in genomic clusters; some but not all of these clusters also include known genes. As EST collections from such tissues and cell types are relatively rare, the fact that EST overlaps with many of the mouse models were found only for these tissues is even more notable. Excluding these unannotated mouse genes, the counts of mouse genes in each ZNF subfamily are roughly similar to those in the human genome; if the novel models are taken into account, the number of KRAB-ZNF genes would be substantially higher in mouse than in the human genome (Table 1).

Identifying "DNA-Binding Orthologs" for Human ZNF Genes

To identify ZNF genes encoding proteins with conserved DNA binding preferences, we extracted the ZNF DNA-contacting residues from translated gene models, including the dog and curated human gene models from our previous study[66]. We then carried out a global alignment of these fingerprint sequences from all species (see Methods). After an initial alignment, we found a number of deeply conserved protein groups (e.g., conserved fingerprints in mouse and non-mammalian species) that did not include a human protein member. Most of these cases involved known human genes encoding only one or two ZNF tandem domains in any single exon; these genes would have been missed with our previous approach. To include the missing human proteins in this analysis, we collected the human protein sequences from GenBank, extracted fingerprint patterns, and repeated the global alignments for a final set (Appendix A Table A2).

In addition to fingerprint alignments, we used reciprocal best Blast, a standard method for ortholog identification used in most published studies[66,87,88]. Reciprocal Blast was the only way to positively identify orthologs in many large groups, like the SP1 and KLF families, which include large numbers of proteins with

identical fingerprints (Appendix A Table A2). Fingerprint alignments also clustered together groups of paralogs with similar fingers including lineage-specific duplicates; fingerprint alignments could not always resolve these groups.

We consolidated and manually curated the results from Blast and fingerprint matches to identify groups of orthologs to the human protein set (Appendix A Table A3). Using these combined data, each human protein was classified as 1) primate-specific (detected in human only); 2) shared by eutherians (human and dog and/or mouse); 3) shared by mammals (at least human and opossum); 4) shared by amniotes (at least human and one bird or lizard); and 5) shared by tetrapods (at least human and frog) (Table 4). Of special note, nine of the unannotated mouse models we discovered (discussed earlier) encode predicted proteins with fingerprint patterns that match annotated genes in human and other species extremely well. For example, four of the novel mouse models are clearly conserved in dog and human, and one model, matching human gene ZNF853, detects clear orthologs in dog, opossum, and chick (Tables 2 and 3). We counted genes (including PRDM9) with excellent, unique best-Blast matches in other species but no fingerprint match, as conserved genes with divergent ZNFs.

Here, we should note that these classifications should be considered as a minimal depth of conservation, as orthologs might be found by scanning additional species, different evolutionary groups, or finished genomes as they become available. As an example, several of the human genes conserved in frog, such as ZBTB16 (aka PLZF1), also recognize orthologs in Drosophila. Nevertheless, the classifications provide a solid overall view of family and subfamily history in vertebrate lineages.

As summarized in Table 4 and consistent with previous estimates[66] , the KRAB-ZNF family contributes the vast majority of ZNF genes that are exclusive to eutherians or to primate lineages. In contrast, nearly all the human genes that are functionally conserved across amniotes or tetrapods are members of the ZNF-only and BTB/POZ-ZNF subfamilies. We also found SCAN-ZNF and KRAB-ZNF genes in most species, although no SCAN-ZNF and very few KRAB-ZNF proteins were conserved across vertebrate groups. Findings from each subfamily are highlighted further in following sections.

A Small Number of Deep Vertebrate Roots for the Human KRAB-ZNF Family

Of the 366 human protein-coding KRAB-ZNF (not including the SCAN-KRAB-ZNF genes, which are discussed later), only 181 genes (49.5%) found a convincing and unique (1:1) fingerprint match in one or both of the

other eutherians; 185 genes were classified as primate-specific (Table 4). Only 17 human KRAB-ZNF genes were found with fingerprint sequence conserved between eutherians and opossum. Looking for orthologs in non-mammalian species, we found only three human KRAB-ZNF proteins, ZNF282, ZNF777, and ZNF783, that have orthologous proteins in non-mammalian amniote groups; orthology is recognized both by overall protein sequence and fingerprint pattern similarities. In particular, the fingerprints of human, bird, and lizard ZNF777 and ZNF282 proteins are strikingly similar, as illustrated by the alignment of ZNF282 orthologs (Table 5). In contrast, while the lizard ortholog of ZNF783 is clearly similar in overall protein sequence and was identified as the best match to mammalian ZNF783 in fingerprint alignments as well, two ZNF motifs are deleted in our lizard gene model compared with the mammalian orthologs (Appendix A Table A2).

Notably, ZNF282, ZNF777, and ZNF783 are clustered as neighbors in the distal end of human chromosome 7 (chr7; cytogenetic band 7q36.1; Figure 1). These three genes and their cluster neighbors, ZNF398, ZNF212, ZNF746, and ZNF767, correspond to 7 of the 17 total KRAB-ZNF genes that are conserved between human and opossum. The orthologous opossum genes are also clustered in chr8 and although the bird and lizard genomes are mostly too fragmented to assess clustering, ZNF777 and ZNF783 are also found clustered in zebra finch chr2 (Figure 1).

We predicted 158 intact KRAB-ZNF genes in the frog genome including 112 that overlap with ENSEMBL gene models (Table 1 and Appendix A Table A1). However, none of these or other predicted frog ZNF genes had a significant fingerprint match with a mammalian KRAB-ZNF gene. To test another measure of relatedness, we examined KRAB-A sequence similarity, an approach we have used successfully to assess KRAB-ZNF subfamily relationships in the past[66,89]. We aligned all ZNF-associated KRAB-A sequences from human, opossum, chicken, and frog and created a maximum likelihood tree of these sequences. We rooted this diverse collection of KRAB-A domains on the KRAB domain of the sea urchin (Strongylocentrotus purpuratus) PRDM9 protein to gain a more global view of lineage-specific expansions and distant relationships between subfamily groups (Figure 2 left).

The tree yielded several interesting results that together shed new light on the early history of KRAB-ZNF subfamily. In particular, after a branch leading to human PRDM9 and its primate-specific paralog, PRDM7, the second branch to emerge from the S. purpuratus PRDM9 root includes KRAB domains from human ZNF282, ZNF777, ZNF783, and other members of human chr7q36.1 cluster (Figure 2 right). Curiously, this clade also

includes the KRAB domain of ZNF862, which contains TTF-type fingers rather than the C2H2 type and is unrelated to the ZNF282 family but nevertheless clusters with them in human, other eutherians, and in the opossum genomes (not shown).

KRAB domains from different species were otherwise largely segregated in the tree with only a few clades including sequences from more than one lineage. As expected, one large clade mostly comprised human KRAB sequences (red bars in Figure 2 left), but we also identified one large clade comprised only of genes from opossum (429 opossum genes; purple bars) and another isolated clade from frog (142 genes; green bars). These groups suggest that the KRABZNF genes we observed in those species are derived from lineage-specific expansions in amphibians and marsupials that are very similar to those that have been documented in detail for eutherians.

The KRAB A domain has diverged significantly from the PRDM9 root in all species examined. But the PRDM9 (not shown) and ZNF282-related KRAB domains[72] lack the amino acid sequences that are known to be essential for KAP1 interaction. Specifically, five amino acids, conserved in two clusters within canonical mammalian KRAB A domains (DV at positions 6,7 in the human consensus in Figure 3 and MLE in positions 36-38), have been shown to be essential for KAP1 binding[67,90]. If all KRAB A domains from each species are assembled into a consensus sequence, it is clear that the C-terminal MLE cluster is also absent from the majority of frog KRAB A sequences (Figure 3). The opossum consensus includes both clusters of KAP-binding amino acid sequences spaced similar to the canonical human domain; these KRAB sequences are thus likely capable of binding KAP1. Chicken genes also include conserved amino acids in most positions although spacing between the two clusters is relatively condensed compared with the human consensus, and the lizard consensus sequence lacks the central leucine in the essential MLE cluster (Figure 3). The avian and lizard KRAB domains may interact with the KAP1 corepressor, although this function cannot be assumed without experimental testing. The KRAB_A domains of genes in human 7q36.1 cluster, which are closely related to ZNF282, also lack the ‘MLE’ binding site (Figure 3).

SCAN- and SCAN-KRAB-ZNF Subfamilies

The SCAN domain was exapted from a Gypsy retrotransposon element and incorporated into ZNF-containing gene structures in tetrapods, most likely during or just preceding the emergence of amniotes[91]. Although we predict a single frog SCAN-ZNF gene (ZF02611) which overlaps well with a X. tropicalis ENSEMBL model,

ENSXETT00000023617, this gene was interpreted by the previous authors as a misassembly or erroneous gene prediction, and without experimental evidence we cannot comment further on its validity. The genome of the lizard, A. carolinensis, was shown previously to include many polydactyl ZNF genes that include either SCAN or an ancestral version of the domain[91], and we did predict more than 300 SCAN-containing ZNF genes in this species (Table 1 and Appendix A Table A1). However, none of these lizard genes were identified, either by reciprocal Blast or by fingerprint matches, as convincing candidate orthologs for ZNF genes of any type in any of the mammalian groups. This suggests that like the KRAB-ZNF subtype in frogs and other species, SCAN-ZNF loci expanded independently in reptilian lineages.

BTB/POZ-Containing and ZNF-Only Genes

The majority of BTB/POZ-ZNF genes are quite ancient, with orthologs in most or all species examined (Table 1 and Appendix A Table A1). Four BTB/POZ-ZNF gene copies appear to have been acquired as novel genes in amniotes. However, alignments suggest that genes in this subfamily have evolved under some pressure for ZNF divergence (Appendix A Table A2 and discussed later). For example, one gene that would be counted as "primate-specific" based on fingerprint matches alone, ZBTB48, is actually well conserved in overall protein sequence and syntenic location in amniotes, but proteins from the different species bear non-recognizable similarities in DNA-contacting amino acid residues. Indeed, based on reciprocal best-Blast and conserved genome locations, none of the human BTB/POZ-ZNF genes appear to be specific to primates, although several of the genes including ZBTB48, ZBTB41, ZBTB44, and ZBTB49 display highly diverged fingerprint sequences across the different vertebrate lineages (Appendix A Table A2). Most BTB/POZ-ZNF genes we identified have only two or very few fingers, and mutations in one finger can thus have a dramatic effect on fingerprint similarity scores and presumably, overall protein function.

The 212 human "ZNF-only" models we counted are distributed throughout primate-specific, eutherian, mammalian, amniote, and tetrapod evolutionary groups, with the largest number of genes showing evidence of tetrapod (or earlier) origins (Table 4). This group, together with the BTB/POZ genes, includes most of the deeply conserved polydactyl ZNF genes.

Fingerprints in Many Orthologous ZNF Groups Are Evolutionarily Divergent

Combining Blast and fingerprint alignment and focusing particularly on human-mouse comparisons, we identified a selection of genes with strong Blast identities but significant fingerprint diversity, and also a few genes with the opposite pattern (Appendix A Table A3). One example of an ancient, highly conserved gene encoding a protein with a divergent fingerprint pattern is ZNF507, a gene that has recently been implicated as a novel risk factor for human neurodevelopmental disorders[92]. The ZNF507 fingerprint patterns suggest a complex history, with certain ZNF positions having been selectively deleted or diverged through missense mutations in certain lineages, while retained strictly in order and sequence in other evolutionary groups (Table 5). For example, human and lizard retain the exact pattern of four amino acids (TVGN) in ZNF 7 in the alignment, but this ZNF has been lost in other species, including mouse; the mouse protein also differs from human in fingerprint sequence in other ZNF positions. Frog and lizard share sequence in ZNF alignment position 8, suggesting that the motif was ancestral and lost in mammals. A core of three fingers (positions 3, 4, and 5 in the alignment, Table 5) is strictly conserved in this ancient protein for every species examined, suggesting an especially important functional role.

Despite this structural divergence, ZNF507 orthologs have retained very similar patterns of developmental expression, as evidenced by ISH in sectioned midgestation mouse and chicken embryos (Figure 4). Expression of mouse Zfp507 was particularly high at embryonic day 12.5 (E12.5), with intense expression in the developing brain, in the spinal and facial ganglia, and developing facial structures (E12.5 corresponds to Theiler stage [TS] 16). The organ systems of birds and mammals do not develop apace, but we saw remarkable similarities in the pattern of neural expression in TS20 chicken embryos (Figure 4B). The very high levels of neural and craniofacial expression in embryos of these two species fit the predicted neurodevelopmental role of this human gene very well.

As illustrated well by ZNF507, most of the species differences we noted involved the in-phase insertion or deletion (indel) of ZNFmotifs. We also detected groups in which orthologs had similar number and arrangement of ZNFs but divergence in fingerprint sequence, and many cases with a mixture of both types of mutation. These patterns have been noted previously as being common paths to divergence for KRAB-ZNF paralogs and orthologs[66,76,93–96]. However, our alignments show clearly that these same patterns occur frequently in orthologous groups of polydactyl ZNF genes of all types (Appendix A Table A2). We identified only a handful of genes that, like PRDM9[97], vary so dramatically in fingerprint sequence that ZNFs could not be aligned. These

cases include five human KRAB-ZNF genes for which orthologs were detected only in mouse: ZNF160 and mouse Zfp160, ZNF780B/Zfp780B, ZNF658/Zfp658, and ZNF84/Zfp84, and the previously reported pair of ZNF226/Zfp61[94](Appendix A Table A2). Nothing is known about the functions of any of these strikingly divergent genes.

At the other extreme, we found 170 human genes that, like ZNF282 and ZNF777, encode fingerprint patterns that have been rigidly conserved since their inception. This group includes 65 tetrapod-, 29 amniote-, 26 marsupial-, and 50 eutherian-conserved genes (Appendix A Tables A2 and A3). The human genes of this type that are conserved in amniotes or tetrapods include many with well-studied developmental functions, including members of the SP1 and KLF families[98,99]. However, this most highly conserved group also includes many genes for which no functional information is currently available. The rigid conservation of the DBDs in proteins encoded by these genes suggests that they have been selected to maintain essential regulatory roles.

Both Conserved and Primate-Specific Polydactyl ZNF Genes Are Most Highly Expressed in Evolutionarily Divergent Tissues

To gain clues to the functions of the most conserved genes, we examined public gene expression (RNA-seq) data from human adult tissues (Illumina Human Body Map 2.0 or HBM2.0), and including more recent data gained from dissected human term placental tissues[82]. From expression values calculated for uniquely mapped sequence reads by Kim et al. in this published article, we extracted expression data for ZNF genes conserved to tetrapods and amniote species and clustered the data to view gene expression patterns as heat maps. Expression patterns vary significantly over this group, but clusters of genes showed similar expression with highest mRNA levels in 1) lymph nodes and white blood cells, 2) ovary, prostate, and testis or 3) placenta.

Interestingly, among the genes expressed at highest levels in placenta are the most ancient members of the KRAB-ZNF family: ZNF777 and ZNF282 (the two genes cluster as indicated by the arrow in Figure 5A).

For comparison, we also examined expression patterns of the primate-specific polydactyl ZNF genes in the same RNAseq data set (Figure 5B). These recently duplicated genes also displayed enrichment for expression in immune and reproductive tissues, with ovary being the most common site of highest expression. Expression patterns for the conserved and the primate-specific gene sets were thus very similar, although primate-specific genes are

relatively more enriched in adrenal gland and relative few primate-specific genes are expressed highly in skeletal muscle or in the amniotic and chorionic components of the placenta (Figure 5).

For further information regarding the functions of ZNF282 and ZNF777, we carried out two sets of additional experiments. First, we used quantitative RT-PCR (qRT-PCR) to measure expression levels in RNA isolated from dissected mouse embryos and placenta collected at successive days from embryonic day 12.5 (E12.5) to E18.5 (which is just before birth in mice). The transcripts were expressed with distinct patterns in the dissected decidua, fetal placenta, yolk sac, fetal head, fetal body, and fetal liver of mouse embryos across midgestation development (Figure 6A and B; tissues presented in the order listed above for each gestational stage). Placenta expression for both ZNF genes were high in both fetal and maternal components and fetal membranes at the latest gestational stages (E18.5), concordant with the high levels of expression seen in human term placenta (Figure 5A).

To extend expression analysis for these two genes to the level of cell type in placenta, we performed IHC experiments using commercial antibodies in paraffin-embedded sectioned human tissues. Concordant with RNA-seq experiments, ZNF777 and ZNF282 are both highly expressed in multiple cell types in both fetal and maternal components of the human term placenta (Figure 6C and D). More specifically, high levels of expression for both proteins were detected in decidual cells in thematernal compartment (Figure 6E) and in the syncytiotrophoblast cells lining the anchoring and floating chorionic villi (Figure 6F and G). Unlike ZNF777, ZNF282 is also very highly expressed in a subset of lymphocytes within the maternal blood spaces (lc in Figure 6D).

Discussion

With their predicted involvement in transcriptional regulation and their unusually dynamic evolutionary histories, vertebrate polydactyl ZNF genes have commanded a substantial amount of analytical attention. However, despite these efforts, much about their evolution and function has remained unclear. This includes their family history, their patterns of conservation or non-conservation, and specifically their orthology relationships and functional similarities across species. The present study adds new clarity to the ZNF family picture in several respects. First, by creating gene models de novo we were able to compare not only established gene models as several other studies have done[88,100,101] but also to include recent lineage-specific pseudogenes and novel, especially unannotated protein-coding genes. These include at least 122 novel mouse gene models, all of which are

supported by EST evidence; intriguingly, the supporting ESTs are overwhelmingly derived from early embryonic sources. The expression of these novel mouse genes in early embryos would fit well with recent data tying known polydactyl ZNF genes to cell fate decision making and early development[88,102–106] . Extrapolating this information to other species, it seems likely that many more of the novel models we found will represent functional, developmentally active genes.

Second, by examining both Blast-based and DNA-binding amino acid fingerprint similarities, we identified clear cases of orthologous groups in which there has nonetheless been significant divergence in fingerprint sequence over evolutionary time. In a small number of cases, the exemplar of which is PRDM9, we found clearly orthologous genes with no discernible fingerprint similarity across species. However, fingerprint divergence for most orthologous groups involves the in-phase deletion or tandem duplications within the ZNF array, similar to the pattern we and others have noted for KRAB-ZNF orthologs in the past [66,71,93,95]. Neuronally expressed ZNF-only gene, ZNF507, provides an excellent example (Table 5). The ZNF507 fingerprint alignments could suggest that the protein has evolved to favor different DNA-binding motifs in each species; alternatively (or perhaps additionally), they could point plainly to three conserved ZNF motifs as having the most essential regulatory roles. In either case, the alignments we present may provide a useful resource as members of this large TF family are targeted for functional characterization.

This ZNF “indel” pattern of divergence was observed for all subfamilies of polydactyl ZNF genes, even within orthologous groups (like ZNF507) that are otherwise relatively well conserved. It thus seems likely that it is the tandem arrangement of ZNF-encoding motifs, per se, that confers a propensity for ZNF indel generation, possibly through a replication slippage mechanism[95]. If this model is correct, strong selection pressure would be required to maintain rigid conservation of the number and order of motifs within ZNF arrays. It is therefore especially noteworthy that the ZNF motifs in hundreds of genes have been strictly conserved in number and sequence over millions of years of vertebrate evolution.

The group of genes showing this highly conserved pattern is dominated by ZNF-only and BTB/POZ-ZNF gene subtypes, including many that are known to regulate critical steps in differentiation and development[99,107–109] .Intriguingly, however, many unstudied genes, including members of the exceptionally dynamic KRAB-ZNF subfamily, are also highly conserved. The most ancient of these conserved human KRAB-ZNF genes, ZNF282 and

ZNF777, a more diverged but ancient relative, ZNF783, and more recently derived cluster neighbors stand out in mammals for their inclusion of an unusually structured KRAB domain that does not bind KAP1 and functions as a transcriptional activator[72,73]. Evolutionary analysis supports the ancient provenance of this activating KRAB and reveals that KRAB-ZNF genes with similar KRAB domains have expanded independently in amphibians and reptiles. The canonically structured, KAP1-binding repressive version of the KRAB A domain is dominant only in mammals, although a similar (and possibly still KAP1-binding) sequence is also the dominant version of the ZNF-linked KRAB A domain in birds.

The dramatic expansion and rapid divergence of repressive KRAB-ZNF genes in mammals has suggested their participation in an “arms race” with the need to silence endogenous retroviruses (ERVs) hypothesized as the dominant driver (Thomas and Schneider 2011). Supporting this notion, a handful of KRAB-ZNF genes have been shown to silence retroviral sequences by binding to motifs within their flanking long terminal repeats (LTRs). Intriguingly, human ZNF282 (also called HUB-1) is one of this very small number of verified LTR-binding KRAB-ZNF proteins, recognizing a motif within the U5RE regulatory region of the human T-cell leukemia virus (HTLV-I) LTR and repressing viral activity. Although the KRAB domain in ZNF282 and other cluster relatives activates transcription, these proteins also include a second domain, called HUR, which confers repressive activity. Rather than acting simply as HTLV-I inhibitor, ZNF282 has been proposed to facilitate an alternative path for the virus by promoting latent infection[72]. Interestingly, one cluster relative, ZNF398, can generate HUB-containing repressive or HUB-minus activating isoforms through alternative splicing[73], and ZNF282 may be able to do the same. Thus, ZNF282 may have evolved to regulate retroviral sequences but has a complex relationship with the virus that cannot be described in simple arms race terms. Indeed, it may be possible that pre-established ZNF282 binding motif, which evolved for other purposes, was captured and domesticated by HTLV-I. Genome-wide DNA binding assays in cells and tissues of different species should allow us to dissect the history of this intriguing interaction.

Whatever the model, the rigid conservation of fingerprint patterns in polydactyl ZNF proteins suggests that their DNA binding activities have evolved essential biological roles. Indeed, naturally occurring or targeted mutations in many deeply conserved polydactyl ZNF genes confirm essential roles in differentiation and development in humans and model organisms[99,107–110]. We hypothesize that other coexpressed genes in this

highly conserved cohort are also associated with unstudied and important developmental functions, albeit functions that in many cases may be challenging to discern. For example, in light of their antiquity and very tight DNA-binding conservation, the high expression of ZNF777 and ZNF282 in placenta—within cell types that vary significantly even between humans and mice including some that do not exist in lizards and birds—is particularly puzzling. Although these placental cells are lineage-unique, they evolved from fetal membranes and uterine cell types that are common to all amniotes [86,111,112]. However, these cell types and structures have continued to evolve independently in every species, making placenta the most evolutionarily divergent of all mammalian tissue types[95].

In fact, the rapid pace of placental divergence reflects another type of arms race—that between the interests of the mother and developing fetus— which is a defining feature of mammalian biology[113] . Similar types of evolutionary battles have played major roles in shaping vertebrate reproductive tissues and cell types, with wide impact on species-specific morphology, metabolism, and behavior[112,114]. Given the very high levels at which both conserved and primate-specific polydactyl ZNF genes are expressed in reproductive tissues, we propose that this larger evolutionary arms race has been the real driver of polydactyl ZNF expansion and divergence in vertebrate history. The facility with which polydactyl ZNF genes can diverge to generate opportunities for DNA-binding diversity makes them ideal raw materials for crafting novelty in gene regulatory pathways. The data provided here identify prime targets, in the form of deeply conserved and unique genes and proteins, for testing these hypotheses in future studies.

Chapter 3. Conservation Across The Metazoa Of Differentially Expressed Genes Associated With Honeybee Behavioral Phenotypes[2]

Abstract

The emerging field of sociogenomics explores the relations between social behavior and genome structure and function. An important question within sociogenomics is the extent to which associations between social behavior and gene expression are conserved among the metazoa. Prior experimental work revealed differential brain gene expression patterns between African and European honeybees, and within European honeybees with different behavioral phenotypes. The present work is a computational study of the above-mentioned work in which we analyze by orthology determination, the extent to which the differentially expressed genes are conserved across the metazoa. We go on to employ an alignment-free similarity measure on the promoter regions of human orthologs to infer the existence of human networks orthologous to the honeybee networks. We find that the differentially expressed gene sets associated with alarm pheromone response, with the difference between old and young bees, and with the colony influence on soldier bees, are enriched in widely conserved genes compared to the honeybee genome overall, indicating that these differences have genomic bases shared with many other metazoans. On the other hand, the sets of differentially expressed genes associated with the differences between African and European forager and guard classes are depleted in widely conserved genes, indicating that these differences are genomically relatively specific to honeybees. For the alarm pheromone response, a particularly high degree of conservation is with the mammals—even higher than with other insects. Gene Ontology identification of the human orthologs to the strongly conserved honeybee genes associated with the alarm pheromone response shows strong representation in functional groups associated with protein metabolism, regulation of protein complex formation, and protein folding.

Introduction

[2] D2z analysis and related discussion are credited to Miriam Ruth Kantorovitz. Honeybee data were generated by Professor Gene E. Robinson's laboratory. Initial outline of the project was generated by Professor Eric Jakobsson. Professor Eric Jakobsson and Professor Gene E. Robinson contributed continuing discussion during the course of the project.

Social behavior is defined as behaviors that involve interactions between multiple entities of the same species which influence immediate and future behaviors[115]. There exist strong connection between gene, brain and social behavior. Various brain regions related to social behavior regulation in mammals have been discovered[116]. Social information is able to alter gene expression and therefore induce different behavioral phenotypes. For example, the social-behavior alteration of egr1 expression has been seen in zebra finch[117], cichlid fish[118] and rat[119]. Although specific behavioral outcomes vary widely from species to species, the biological needs that drive these behaviors and even the genes underlined can be highly conserved[120]. And it is believed that there are conserved genomic bases that control the social behavior[115]. However it is only recently that it became foreseeable for sufficient genomic data and techniques to become available to explore the genomic correlates of social behavior[121]. In addition to our own human species, our two nearest relatives, the chimpanzee and the bonobo, have been sequenced[122,123]. The genomes of these three species are almost equidistant from each other in similarity and much closer to each other than to any other species, leading one author to refer to humans as "The Third Chimpanzee" [124]. The differences and similarities in social behaviors of these three species bear on the most significant behavioral issues facing us, such as violence, altruism, care of young, etc. Presumably these differences and similarities have correlates in the differences and similarities in our genomes. However to reliably connect genotype to behavioral phenotype across our species would require experiments completely unethical to perform on humans or human-like animals.

Thus non-primate model organisms are needed to understand the genomic basis of social behavior. Based on behavior patterns, the honey bee seems a very appropriate model organism. Experiments to link gene expression patterns to social behavioral characteristics and environmental stimuli are feasible, and the honey bee genome has been completely sequenced[125]. In addition, different members have well-defined social roles in the life of the honey bee colony. The entire social organization of the colony is based on the genome of the individual honey bee. It is known that the division of labor within the hive is based on genetic differences between individual honey bees and environmental influences that the colony imposes on its members at different stages of the individual's development, on visual, tactile, and chemical signals that the members send to each other, and on influences external to the colony[125]. However, the interplay between these factors is far from defined with respect to variation in particular genes or regulatory domains in the genome. It seems that the individual honey bee's social role is determined by

both the social/environmental factors and the individual's heredity.

While there are compelling analogies between honey bee social behavior and that of humans, it is appropriate to question whether the genomic correlates bear any significant relevance to each other. The last common ancestor of honey bees and humans was 600 million years ago[126]. It could be that the correspondences between honey bee and human social organization and social behavior represent convergent evolution of adaptive behaviors based on completely different genes and genetic circuits. Alternatively, it could be that both honey bee and human behaviors are elaborations and modifications of underlying patterns that were present in a common ancestor of both. In this latter view, some species in the lineages leading to both have lost or inhibited expression of these patterns, while other species such as the honey bee and human continue to express them and use them as a set of building blocks for social behavior. If the latter view is true, comparative genomics of honey bee social behavior and mammalian (including human) social behavior may yield insights into the most fundamental aspects of the genomics of social behavior.

The connection between honey bee sociogenomics on one hand and human sociogenomics on the other must be made by inference of orthology. Unfortunately orthology is of necessity not verifiable in the same fashion as other areas and techniques of bioinformatics, since it involves theoretical reconstruction of an evolutionary history that can't be experimentally replicated. Thus there is no reliable validation set on which to test a method. Different reasonable ways of creating orthologies may give significantly different results[127]. Whether one makes a liberal or conservative interpretation of orthological relationships produced by a particular method depends on the context, in particular whether one is concerned about contamination by false positive identifications of orthologs, or more concerned about loss of information by false negatives (failure to identify real orthologous relationships). In the present paper we make a preliminary test of the hypothesis that the social behavior of honey bees and other metazoans, including humans, have common fundamental genomic building blocks.

Specifically, we utilized the above-cited[50] data sets , which analyzed differential brain gene expression patterns associated with different social classes in both African and European colonies, in bees that were nurtured in both their own communities (European in European colonies, African in African colonies) and in opposite hives (European in African colonies, African in European colonies). African and European honey bees are subspecies of the Western honey bee, Apis mellifera, and they differ from each other in their hive-defense behavior in a number of

ways that have been summarized as a social behavioral counterpart to variations of threshold and intensity of the "flight or fight" response seen in vertebrate organisms; African bees are much more aggressive than European bees [128]. In general different phenotypes may arise from either differences in gene function or from different patterns of gene expression[129]. In the African and European honey bees the genes are so similar that it is presumed that the different phenotypes are largely results of different patterns of gene expression. Based on these data sets to explore the following questions: To what extent are the differentially expressed genes associated with social behavior in the honey bee conserved across the Metazoa? Through analysis of the highly conserved genes, is it possible to infer that there are likely to be gene co-expression patterns associated with social behavior that are common to a wide range of metazoans, including humans? The results of this analysis hold important implications for the specifically anthropocentric question of whether it will be useful to use invasive experiments on animals such as honey bees to gain useful insights into the genomic correlates of social behavior in humans. Our results also will help the utility of the new approach we present here, to use orthology to relate social behavior-related gene expression patterns in an experimental animal to putative expression patterns in other metazoans, including humans.

Methods

Identifying Metazoan Orthologs of Honeybee Genes

First, honey bee genes that showed up on the microarray studied in [50] was selected. This was done based on the annotation file of this Honey Bee Oligonucleotide Microarray. Out of many available methods[127] of defining orthologs, InParanoid[9] was chosen based on extent of coverage of the honey bee proteome and other proteomes of completed genomes in searchable ortholog databases. Then, the second step was, out of all these 'microarray-present' honey bee genes, to find those that are also present in InParanoid. This was done by mapping the BeeBase ID's (which are the ID's used in the data set from[50]) to NCBI Refseq ID's (which are the ID's used in InParanoid for honey bee). 7462 of these 'microarray-present' honey bee genes are present in InParanoid. At the time of the analysis, there were 100 eukaryotic species in InParanoid with 54 of them (including Apis mellifera) being metazoan species. With S.cerevisiae added as a control, the data set used for our analysis had 55 species(*Saccharomyces cerevisiae, Trichoplax adhaerens, Nematostella vectensis, Schistosoma mansoni,*

Pristionchus pacificus, Brugia malayi, Caenorhabditis japonica, Caenorhabditis elegans, Caenorhabditis brenneri, Caenorhabditis remanei, Caenorhabditis briggsae, Lottia gigantean, Capitella spI, Helobdella robusta, Ixodes scapularis, Daphnia pulex, Pediculus humanus subsp. corporis, Acyrthosiphon pisum, Nasonia vitripennis, Apis mellifera,Tribolium castaneum, Bombyx mori, Anopheles gambiae, Aedes aegypti, Culex pipiens, Drosophila grimshawi, Drosophila virilis, Drosophila mojavensis, Drosophila willistoni, Drosophila pseudoobscura pseudoobscura, Drosophila ananassae, Drosophila melanogaster, Strongylocentrotus purpuratus, Ciona savignyi, Ciona intestinalis, Branchiostoma floridae, Tetraodon nigroviridis, Takifugu rubripes, Gasterosteus aculeatus, Oryzias latipes, Danio rerio, Xenopus tropicalis, Gallus gallus, Ornithorhynchus anatinus, Monodelphis domestica, Bos taurus, Equus caballus, Canis familiaris, Cavia porcellus, Rattus norvegicus, Mus musculus, Macaca mulatta, Pongo pygmaeus, Pan troglodytes, Homo sapiens), which we interrogate for orthology with the 7462 InParanoid honey bee proteins.

The Social Behavior-related Gene Sets

Eight sets of social behavior-related differentially expressed genes were used. They are described in outline in Table 6 and in detail in[50]. Detailed information about genes in these sets is shown in Appendix A Table A4.

The Statistics of Ortholog Gene Count

The p-values in Table 7 for the average number of metazoan orthologs for each data set were computed as follows: For each experimental data set, random sets of matching size were sampled from the 7462 honey bee genes that were present in InParanoid database and spotted on the array, and the average number of orthologs per gene was calculated for each random set. This random sampling was repeated one million times and the number of random sets with average ortholog number equal to or larger than the experimental set was counted. The count divided by 10^6 gave us the p-value for the average ortholog number of the test set. Figure 7, 8 graphically shows how the p-values of the average ortholog number of Forager_CG and Alarm_Pheromone sets were calculated. The p-values for the total number of orthologs of each set for each species (Figure 11) were computed similarly.

For calculating the p-value for the CG-WG difference, the KS-test p-values for the CG-WG difference for Soldier, Forager and Guard (Table 16, 0.026, .122, and .612 respectively) were combined using Fisher's

method[130] by R package MADAM (http://r-forge.r-project.org/projects/madam/, last accessed April 4 2014).

For calculating the p-value for over-representation of orthologs of placental mammals(10 species, from B.taurus to H.sapiens) in the Alarm_Pheromone set and over-representation of orthologs of insects(15 species, from P.humanus to D.melanogaster) in the Old_vs_Young set, p-values in each species (Table 14,15) were also combined using Fisher's method.

In addition, we note that the 1631 honey bee genes in the InParanoid dataset that were not spotted on the oligonucleotide microarray in[50] (and therefore excluded for our analysis), were much less conserved than the genes spotted on the microarray. The distribution of conservation of the honey bee genes spotted on the microarray is given in Figure 9. The y-axis is normalized so the height of each bar represents the fraction of the honey bee genes with the indicated number of orthologs. Just under 5% of the genes had no orthologs in the InParanoid set; within this data set they are unique to the honey bee. Of the 54 species being compared to the honey bee, 12 are insects. The position of the first peak in the distribution (at 15 orthologs) is due to genes that are largely conserved in insects and are uncommon in other metazoan lineages. The position of the second peak (at 50 orthologs) is due to genes that are broadly conserved across the metazoa. Figure 10 shows the degree of conservation of the 1631 genes that were not spotted on the microarray. Approximately a third of those genes are unique to the honey bee. By comparison with Figure 9, it can be seen that the genes that were not spotted on the microarray are much less conserved than those that were spotted on it. This is partly a function of how the microarray was designed[50]. Since the major conclusions of this paper will be based on orthology to other metazoa, and since the genes excluded from the analysis have relatively few such orthologs, the conclusions (especially the p-value results) will not be significantly affected by the exclusion of these genes.

In presenting and discussing the results, we use the term "conserved" to be measured by the number of orthologs that a particular sequence has; i.e., the more orthologs a gene or protein has in other species, the more "conserved" the gene is.

Gene Ontology Analysis

Enrichment of the conserved gene sets in particular Gene Ontology categories was determined using the functional annotation tool in the Database for Annotation, Visualization, and Integrated Discovery (DAVID)[131].

All parameters are default except that we use GO_*_TERM 1-5 instead of GO_*_FAT, and we only keep results of p-value <=0.05 instead of 0.1. Extra functional analyses (of various qualities) were also included: OMIM_Disease[132], COG_Ontology[3], SP_PIR_Keywords[133], Up_Seq_Feature[134], BBID[135], BioCarta[136], Kegg_Pathway[137], Interpro Domains[138], Pir_Superfamily[139] and Smart[140].

The raw Gene Ontology results of 'Vertebrate-conserved' and 'Mouse-human-conserved' Alarm_Pheromone genes are listed in Appendix A Tables A5-10 respectively. Figures of Gene Ontology trees(Figures 12-19) are generated by Perl scripts and Cytoscape[141].

Analysis of Promoter Region Similarity to Infer Likelihood of Co-regulation in 'Vertebrate-conserved' Alarm_Pheromone Genes

Our analysis of probability of co-regulation of the human orthologs of important honey bee genes was based on the D2z measure, first described in [51]. The D2z measure is a normalized alignment-free similarity measure that was designed to detect functional similarity of regulatory sequences. The measure is based on the frequencies of common k-words in the sequences. The k-words represent potential binding sites of (unknown) transcription factors (TFs). The rationale is that genes with promoter regions that are highly similar by this measure are more likely to be co-regulated by a common set of TFs than genes whose promoter regions are less similar. Normalization is a significant feature of D2z. The simpler D2 measure of similarity between two sequences is simply the number of common k-words. In the D2z measure, the similarity between two sequences is the number of standard deviations between the number of common k-words actually found and the number to be expected given the background distribution of base pairs in the two sequences. Thus a given number of common k-words between two sequences will receive a higher D2z score if the sequences are very different from each other with respect to statistical distribution of individual symbols in the sequences. The details of computing the D2z score are given in [51] (and can be requested at http://veda.cs.uiuc.edu/cgi-bin/d2z/download.pl, last accessed March 29 2014). It was shown that D2z can accurately discriminate functionally related *cis* regulatory modules (CRMs) from unrelated sequence pairs [51]. In addition, when non-coding sequences of two species of fruit fly were compared, orthologous CRMs had higher D2z score than orthologous non-CRM sequences. A version of the D2z measure was used in [142] to discover novel CRMs, by scanning the genome for sequences similar to a set of known modules. The discoveries

were validated experimentally in two model systems: fruit fly and mouse.

More recently [143] it was shown that D2z, applied to promoter region of genes, can be used to detect co-regulatory relationship between genes implicated by GWAS in childhood cognitive ability (g), as well as finding co-regulatory relations between the GWAS genes and other genes that are known to be involved in intellectual (dis)abilities.

In this work we use the method developed in [143] to examine the regulatory relationship between human orthologs of highly conserved honey bee behavioral genes, and to find additional human genes that may be involved in the same regulatory co-expression patterns.

For this analysis we considered the human orthologs of the vertebrate conserved genes in each data set. These are a subset of the set consisting of genes that have orthologs in all vertebrate species of InParanoid database (analysis on orthologs of genes only conserved in mouse and human is also done, see Figure 23-25). We excluded genes that were not in the Transcription Start Site database DBTSS version 8 [144]. The resulting list of human genes used for this analysis can be seen in Appendix A Table A12. These human genes were used as "probes." The regions probed were 1000 base pairs upstream and 200 base pairs downstream of the transcription start site in all the human genes in the DBTSS database, using the D2z similarity measure [51]. For D2z, the size of the k-words was taken to be 5, which is approximately the core length of TF binding sites.

Results

Ortholog Distribution Across the Metazoa of Honeybee Genes Related to Behavioral Phenotypes

Table 7 provides the overall summary of the results. At the .05 significance level three of the sets are selectively enriched in genes conserved across the Metazoa: the Alarm_Pheromone set, the Old_vs_Young, and the Soldier_CG set. By the same standard of significance, the Guard_CG, Guard_WG, and the Forager_WG sets are significantly depleted in highly conserved genes.

Figure 11 shows the conservation pattern broken down to each of the 54 species used in the analysis, via a heat map, for the eight data sets of Table 6. Figure 11 reveals details about the clade distribution of conservation that

are not visible in Table 7. It is seen in both Table 7 and Figure 11 that a relatively high degree of conservation is distributed across a wide range of metazoans for Old_vs_Young, Alarm_Pheromone, and Soldier_CG sets. For Soldier_CG and Old_vs_Young, the most significant conservation (red color) is within the insect group. For the Alarm_Pheromone set, on the other hand, the most significant conservation (red color) is in placental mammals. This figure suggests that, of all the gene sets analyzed, the set that is differentially expressed in response to the alarm pheromone stimulus is especially promising with respect to being involved in genetic circuits common to honey bees and mammals. In order to be conservative in our assignment of orthologs (minimize false positives, even at the expense of incurring false negatives) we chose for detailed further analysis the set of 85 genes that are differentially expressed in the alarm pheromone response and conserved in ALL the vertebrate species (altogether 19 vertebrate species in InParanoid, ranging from T.nigroviridis to H. sapiens) considered in this study. In fact, the p-value for over-representation of orthologs of placental mammals in this set is 9.67e-11 (See Methods), which constitutes a correlation effectively impossible to have occurred by chance. Similarly, the completely colored spaces for all the insect species in the Old_vs_Young set indicate a correlation effectively impossible to have occurred by chance (with a p-value of 0).

A larger set of genes (that are conserved in mouse and human but not necessarily in all 19 vertebrate species) was also analyzed, with results given in supplementary materials. The conclusions are not significantly affected by the differences in the results from these two gene sets.

We note also that in each of the three classes of bees (soldier, forager, guard) where we have both a CG set (differential gene expression between bees raised in predominantly African and European colonies) and a WG set (differential gene expression between genetically African and genetically European honey bee) there is more enrichment in orthologs with other metazoans, in the CG set than in the WG set with a p-value of 0.051(See Methods). This result speaks to the general issue of the interaction between nature and nurture in defining social behavior. It tells us that if we wish to draw inferences for other metazoans from the different behavior of African and European honey bees, we must consider how the colony socializes the individual bees. At the genomic level, this suggests that the nature and extent of genetic diversity within African and European colonies (beyond the scope of the current study) is perhaps more relevant for understanding the broader relevance of the behavior of the different strains than is the difference in the reference genomes for the different strains.

Gene Ontology Analysis of Conserved Alarm Pheromone Genes

We used the DAVID suite of programs to identify the GO categories that are over-represented in the 85 genes mentioned above relative to their overall incidence in the human genome (77 of these 85 genes' human orthologs have Entrez annotations (Appendix A Table A11)), at p-values of 0.01 and 0.05. For better comparison, we performed separate GO analyses for all these 77 genes, genes that were up-regulated (47 of them), and genes that were down-regulated (30 of them). The results are summarized in Figures 12 and 13 (node indexes will be shown in the brackets in the following paragraphs) and in Tables 9-13(raw data are shown in Appendix A Tables A5-7).

Gene Ontology analysis for biological process revealed that the overall the pattern of enrichment in vertebrate orthologs of genes differentially regulated in the honey bee response to alarm pheromone suggests commonalities between the honey bee response network and orthologous patterns in humans in the areas of protein and carbohydrate metabolism, response to stimulus, protein and carbohydrate metabolism, formation of protein complexes and protein folding, and both chromosome and cytoskeleton organization. This is detailed in the following five paragraphs that discuss prominent features of Figure 12.

There is a set of ten strongly enriched GO categories under "cellular component organization or biogenesis" (118). This feature is seen on the left hand side of Figure 12 and in the top section of Table 9. Near the top (less specialized) level these include "cellular component organization" (58), and "cellular component assembly" (23). As we move down this section of the tree to the next level of specialization, we find that the enriched categories deal with macromolecular complexes ("macromolecular complex subunit organization" (95), "macromolecular complex assembly" (26), "cellular macromolecular complex assembly"(82)). And finally at the deepest levels of specialization, we see a focus on proteins with "cellular protein complex assembly" (79), "chaperone-mediated protein complex assembly" (108) and "protein complex assembly" (78). Finally a short parallel branch descending from index 118 is comprised of "cellular component biogenesis" (2) and more specialized "protein complex biogenesis" (56). Taken together this feature of the tree in Figure 12 suggests that the human pattern orthologous to the expression pattern of the honey bee alarm pheromone response involves protein complex organization and biogenesis. And notice that most of the terms are not significant for down-regulated genes (see node face colors in Figure 12).

There is a set of 9 enriched GO categories under "metabolic processes" (16), which is itself enriched. This

feature is seen on the right hand side of Figure 12 and the lower portion of Table 11. Directly under metabolic processes is “primary metabolic processes” (94). Three of the indices in this region of the tree are specifically related to proteins, in particular “protein metabolic processes” (30), “cellular protein metabolic process” (63) and “protein folding” (14). Five of the indices are related specifically to carbohydrates: “monosaccharide catabolic process” (1), “hexose catabolic process” (109), “glycolysis” (3), “hexose metabolic process” (119) and “glucose metabolic process” (13). Taken together these features of this region of the tree suggest that the human pattern orthologous to the expression pattern of the honey bee alarm pheromone response involves modulation of both protein and carbohydrate metabolism.

There is a section of the tree that is comprised of enriched GO categories emanating from “single-organism process” (7). These include “intracellular receptor-mediated signaling pathway” (27), “organelle organization” (104), “cortical cytoskeleton organization” (68), “cortical actin cytoskeleton organization” (38) and “chromatin organization” (53). Taken together these features of the tree suggest that the human pattern orthologous to the expression pattern of the bee alarm pheromone response involves intracellular receptor mediated signaling and some reorganization of both the cytoskeleton and of chromatin in the chromosomes.

There is a section of the tree that includes “children” of the general category “response to stimulus” (67). The enriched more specialized categories under this general category are represented by “response to abiotic stimulus” (62), “response to stress” (87), “response to organic substance” (92), “response to unfolded protein” (83) and “response to biotic stimulus” (12). Taken together these enrichments suggest that the human response pattern orthologous to the honey bee alarm pheromone response also involves responses to chemical and possibly other stimuli. Presumably the response to unfolded protein seen in this section of the tree is related to protein metabolism and biogenesis, and the protein complex assembly that is simultaneously being up-regulated during the overall organism response as indicated in other parts of the tree.

There is a set of three enriched categories lies under the general category of “cellular process” (77). These are “cellular metabolic process” (9), “phosphorylation” (59) and “generation of precursor metabolites and energy” (19). There also are three enriched categories not connected by descent to any other enriched categories, namely: “localization” (111), “establishment of localization in cell” (18) and “regulation of cell cycle” (34).

Gene Ontology analysis for molecular function reveals that that all the enriched GO terms fall under one of two

general categories “binding” (25) and “catalytic activity” (21). Figure 13(left) shows the Gene Ontology tree for molecular function in a manner analogous to how the biological process tree is shown in Figure 12. Similarly, Table 12 provides details about the elements of the tree for molecular function in the same fashion as Tables 9-11 do for the tree in Figure 12. This is detailed in the following three paragraphs that discuss prominent features of the Gene Ontology tree for molecular function.

The most prominent feature of the tree for molecular function is at the right hand side, where the following categories are strongly enriched (<=.01): “nucleotide binding” (42), “purine nucleotide binding” (80), “adenyl nucleotide binding” (31), “ribonucleotide binding” (22), “purine nucleotide binding” (107) and “adenyl ribonucleotide binding” (32). All of these are under a higher level parent “heterocyclic compound binding” (70), under which “nucleoside binding” (39) and “purine nucleoside binding” (110) are also strongly enriched (<=.01). Notice that all these GO terms are not significantly enriched for down-regulated genes.

Another significant feature descends from the “protein binding” (65) category and includes: “kinase binding” (93), “hormone receptor binding” (66), “nuclear hormone receptor binding” (105), “unfolded protein binding” (55), “heat shock protein binding” (86), “binding, bridging” (52),“protein binding, bridging” (85), and “SH3/SH2 adaptor activity” (97).

Enriched categories on the left hand side of this molecular function tree underneath the catalytic activity general category include: “lyase activity” (90) and “aldehyde-lyase activity” (112); “kinase activity” (71) and “protein kinase activity” (8); “phosphotransferase activity—alcohol group as receptor” (4) and “intramolecular transferase activity—phosphotransferase activity” (89).

Gene Ontology analysis for cellular component (Figure 13(right) and Table 13) revealed the enrichment pattern including multiple cell components—cytoplasm, nucleus, mitochondria (only significant for down-regulated genes) and other organelles, and protein and possibly other macromolecular complexes. This might have been expected, since the biological processes and the molecular functions implicated in Figures 12 and 13(left) take place in a variety of cell components.

Since the members of the gene set from which these inferences are derived are conserved across the vertebrates, it is plausible that the inferences are valid for vertebrates in general. However, it should be reiterated that the results described in this section do not refer to the totality of either the honey bee alarm pheromone response nor of a

complete network in humans and other vertebrates. Rather, they refer to components of the honey bee alarm pheromone response network that are widely conserved in vertebrates and have a well-defined Gene Ontology classification in humans. These components were presumably present and possibly part of an interacting network in the last common ancestor of the human and the honey bee about 600 million years ago. Both the honey bee alarm pheromone network and networks in vertebrates that share these components will undoubtedly have other different non-shared components particular to their respective classes of organism.

We also submitted the set of genes to the Ingenuity Pathway Analysis (IPA®) (Appendix A Table A11, Figure 20, 21). The results of this analysis are consistent with the Gene Ontology results presented above. They include:

- 21 out of the 85 genes are involved in the "Post-Translational Modification, Protein Folding, Drug Metabolism" (IPA category) networks (Figure 20).
- These genes are distributed throughout the cell, in both the cytoplasm and the nucleus

GO Analyses using genes conserved in both mouse and human (which is a larger set containing the vertebrate-conserved genes plus others) were also performed (See Appendix A Tables A8-10, and Figure 14-19). Similar patterns to those seen in Figure 12 and 13 were observed.

Analysis of Promoter Region Similarity to Infer Likelihood of Co-regulation in 'Vertebrate-conserved' Alarm_Pheromone Genes

Figure 22 shows the co-regulation graph (derived from D2z promoter analysis) for the human orthologs of the vertebrate-conserved genes in the Alarm_Pheromone set mentioned above ("probe genes"). An edge between two nodes (genes) in the graph means that the two genes are likely to be co-regulated by a similar set of (unknown) transcription factors. We see that the graph is connected, and the number of genes that are co-regulated with at least 2 (or 3) probes is 57 (or 27). This suggests that the genes in our set of human orthologs are likely to be part of a regulatory network. The major hubs in the co-expression patterns are EBF1 (Entrez gene ID 1879) and ZNF521 (Entrez gene ID 25925), which are co-regulated, with 35 and 26 probe genes, respectively. They also are hubs in the co-regulation graph of only up and down genes. Specifically, EBF1 is more related to down-regulated genes since 23 out of 35 probe genes are down-regulated. Both ZNF521 and EBF1 are known to be expressed in the brain and interact with each other [145]. In addition, EBF1 has been associated with conduct disorders in children (including

aggression) [146] and ZNF521 has been suggested to be involved in psychiatric disorders [147]. Other hubs of interests are DRD2 (1813), LSAMP (4045), SDC2 (6383), GAL3ST1 (9514), PIK3R4 (30849) and GPR85 (54329), all of which are expressed in the brain and have been associated with mental disorders, behavior disorders or autism [148–154]. Fifteen other genes are also related to neural disorders (Table 17). Figure 23-25 shows similar co-regulation graphs but with the 186 'mouse-human-conserved' Alarm_Pheromone genes as probes.

Corresponding results for promoter region analysis of the 'vertebrate-conserved' genes in each data set are provided in Appendix A Table A12. The co-regulation graphs for the various datasets (of the human orthologs of the 'vertebrate-conserved' genes) share some major hubs, indicating significant overlap of genes between the datasets. The most connected hub, EBF1 (1879), is common to all the datasets, and its bee ortholog, GB14092 has the highest expression value in the Guard WG dataset (0.51). ZNF521 is a common major hub to most datasets. As stated above, EBF1 and ZNF521 are known to interact with each other and are implicated in a range of behavioral and psychiatric conditions in humans.

Gene Ontology analysis of all the hubs from all the datasets indicates that the "Conduct disorder and ADHD" term in the OMIM disease category is enriched (data not shown). All the co-regulation graphs except "GUARD WG" (which is also one of the smallest sets) are well connected. Thus even though some of the datasets of the honey bee co-expressed genes are not enriched in 'vertebrate-conserved' genes, the human orthologs of the 'vertebrate-conserved' genes in these datasets are projected by D2z to be co-regulated. This suggests the existence of orthologous circuits containing the orthologous genes.

Discussion

This work demonstrates the utility of using orthology relationships to study the extent to which the differential gene expression patterns in a model organism can suggest the existence of co-expression patterns that are widely shared across metazoans, including humans. Our analysis supports the following conclusions:

- The collection of genes associated with the honey bee brain response to alarm pheromone is strongly enriched in genes conserved in mammals, and somewhat less conserved in other metazoans. This result suggests conservation of underlying gene networks for social behavior.
- GO analysis of the human orthologs of the differentially expressed Alarm_Pheromone genes shows

strong enrichment in genes associated with response to both biotic and abiotic stimuli and intracellular signaling. One overall theme of the enriched categories is that brain cells are responding to a social stimulus by utilizing carbohydrate metabolism to support significant restructuring of intracellular signaling pathways in ways that involve both the nucleus and the cytoplasm. A second theme relates to protein folding and metabolism. This discovery provides a provocative link between changes that occur during normal bee behavior and neurodegenerative disease, which often involves defects in protein folding [45].

- The differential brain expression patterns associated with soldier bees as a function of the colony in which they are raised is strongly conserved across the Metazoa. A similar but less extreme pattern pertains also to the forager and guard groups. These results suggest that the social effects on brain gene expression and behavioral responses in honey bees involve mechanisms that are common across the Metazoa.
- Guard differential brain gene expression based on individual genotype is strongly conserved across the insects but not across the vertebrates and nematodes, suggesting that the underlying mechanisms are more insect-specific.
- Analysis conducted by the D2z method of promoter regions for the human orthologs of the conserved differentially expressed Alarm_Pheromone genes revealed strong similarities in inferred transcriptional regulation. This result suggests strong similarities in socially responsive genetic circuits common to honey bees and mammals. This may include human transcriptional regulatory networks associated with the OMIM disease category “Conduct Disorder and ADHD”.

Our hope is that future work that elaborates on the approach presented in this paper will be useful for elucidating the molecular evolution of social behavior and for applying that understanding to the relationship between genomic and environmental influences on social behavior in both humans and other animals.

Chapter 4. Future Perspectives

The ‘fingerprint’ alignment method initiated here has helped us to discover novel and interesting evolutionary patterns about zinc finger genes. It perfectly conforms to the general principle of domain-based approaches. Although there have been arguments about functional orthologs previously[3], domain-based orthology discovery is still not as prevalent as conventional nucleotide-based or residue-based approaches such as BLAST. Nevertheless, there have been cases whereby a domain-based approach successfully discovered the functional ortholog where BLAST fails. For example, it was used for discovering an entire group of prokaryotic orthologs to eukaryotic pentameric ligand-gated channels[155]. Such an approach has been further applied and elaborated for bacterial and archaeal homolog discovery. And a combination of different similarity measures based on domain-decomposition and the frequencies of each domain have been used[156]. However, the specific domain evolutionary events, like insertion and deletion, are not highly visible from the similarity scores alone. Therefore, it would be of great interest to combine the alignment approach formulated in this dissertation with the similarity measures developed previously[156]. This would become a method for identifying functional orthologs based on domain compositions (or domain sequences). This idea can be generalized as a pipeline as in Figure 27: by integrating InterProScan[40], HMMER(http://hmmer.janelia.org/), MEME/MAST[157,158] and the motif alignment, the pipeline would be able to do a comparative analysis at the domain-level, which could be readily used for domain-based orthology identification. The resulting pair-wise score matrix could be directly exported to other distance-based analysis methods such as clustering. And the resulting multiple coarse grained alignments (aligning domains rather than individual residues or bases) could be visually analyzed for gene comparison. A similar pipeline called “MotifNetwork”[159] has been described but not released for distribution.

On the other hand, instead of being used to define orthologous genes, the domain approach might be extended to the issue of ortholog distribution among species, which is central to chapter 3 of my dissertation. Currently orthology takes an “all-or-none” approach: i.e., either a gene in one organism is orthologous to a gene in another organism, or it is not. The fact that domains/motifs, such as zinc fingers, are inserted and subsequently descended as blocks within genes could lead to the introduction of the concept of “degree of orthology” in which two genes might share some domains, but not others. This approach might help to resolve the current situation, in which different

methods of determining orthology give significantly different results.

Moreover, as the “domain sequence” is relatively short compared to a complete gene or protein sequence, an optimal multiple alignment[160,161] will be feasible for small sets of sequences (which will definitely improves the alignment quality): the optimal multiple alignment of 6 sequences with 10 domains each takes about 50 minutes on a single core of Intel Core i7 2.4 GHz. When there are a lot of motif sequences, the optimal multiple alignment algorithm can be applied hierarchically according to the guide tree (see Figure 26 as a demonstrating example) and alleviate the inaccuracy of a guide tree. If multi-cores are available, the running time for aligning more sequences can be further improved by aligning independent sub-branches of the guide-tree in parallel.

In addition, the improvement of ortholog data sources[26], new data about social-animal gene expressions and the improvement of genome annotations[162] might also greatly aid us toward a more precise pattern of the ortholog distribution of social-behavior related genes across vertebrates. Moreover, it is still unclear as to what kind of functions these genes have in the regulation of social-behavior and whether the underlying regulatory networks are conserved across vertebrates or not. It is found in Chapter 3, for example, that ‘protein folding’-related genes are relatively enriched as conserved genes related to one type of social behavior. Along with evidences about protein-folding related neuron diseases, it would be interesting to explore why protein-folding is so important for behavior-regulation. Additionally, the big task is to move from orthologous genes to orthologous networks and pathways. Studies about the structure of the regulatory network in honeybee has already begun[163], which is based on regression and correlation of gene expression data. By combining sequence-information based analysis like D2z[51] with methods based on expression data, the accuracy and scope of our estimation of the regulatory network might be improved.

Bibliography

1. Pagani I, Liolios K, Jansson J, Chen I-MA, Smirnova T, et al. (2012) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 40: D571–9. Available: http://nar.oxfordjournals.org/content/40/D1/D571.long. Accessed 20 January 2014.

2. Alföldi J, Lindblad-Toh K (2013) Comparative genomics as a tool to understand evolution and disease. Genome Res 23: 1063–1068. Available: http://genome.cshlp.org/content/23/7/1063.long. Accessed 26 February 2014.

3. Tatusov RL, Koonin E V, Lipman DJ (1997) A genomic perspective on protein families. Science 278: 631–637. Available: http://www.ncbi.nlm.nih.gov/pubmed/9381173. Accessed 25 September 2013.

4. Koonin E V (2005) Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39: 309–338. Available: http://www.annualreviews.org/doi/abs/10.1146/annurev.genet.39.073003.114725. Accessed 20 March 2014.

5. Kristensen DM, Kannan L, Coleman MK, Wolf YI, Sorokin A, et al. (2010) A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. Bioinformatics 26: 1481–1487. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2881409&tool=pmcentrez&rendertype=abstract. Accessed 31 March 2014.

6. Kristensen DM, Waller AS, Yamada T, Bork P, Mushegian AR, et al. (2013) Orthologous gene clusters and taxon signature genes for viruses of prokaryotes. J Bacteriol 195: 941–950. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3571318&tool=pmcentrez&rendertype=abstract. Accessed 19 March 2014.

7. Sonnhammer EL., Koonin E V (2002) Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet 18: 619–620. Available: http://www.sciencedirect.com/science/article/pii/S0168952502027932. Accessed 31 March 2014.

8. Berglund A-C, Sjölund E, Ostlund G, Sonnhammer ELL (2008) InParanoid 6: eukaryotic ortholog clusters with inparalogs. Nucleic Acids Res 36: D263–6. Available: http://nar.oxfordjournals.org/content/36/suppl_1/D263.short. Accessed 28 March 2014.

9. Östlund G, Schmitt T, Forslund K, Köstler T, Messina DN, et al. (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res 38: D196–D203.

10. Alexeyenko A, Tamas I, Liu G, Sonnhammer ELL (2006) Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics 22: e9–15. Available: http://bioinformatics.oxfordjournals.org/content/22/14/e9.short. Accessed 27 March 2014.

11. Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178–2189. Available: http://genome.cshlp.org/content/13/9/2178.short. Accessed 19 March 2014.

12. Schneider A, Dessimoz C, Gonnet GH (2007) OMA Browser--exploring orthologous relations across 352 complete genomes. Bioinformatics 23: 2180–2182. Available: http://bioinformatics.oxfordjournals.org/content/23/16/2180.short. Accessed 27 March 2014.

13. Goodman M, Czelusniak J, Moore GW, Romero-Herrera AE, Matsuda G (1979) Fitting the Gene Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences. Syst Biol 28: 132–163. Available: http://sysbio.oxfordjournals.org/content/28/2/132.short. Accessed 27 March 2014.

14. Page RD, Charleston MA (1997) From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. Mol Phylogenet Evol 7: 231–240. Available: http://www.sciencedirect.com/science/article/pii/S1055790396903905. Accessed 27 March 2014.

15. MIRKIN B, MUCHNIK I, SMITH TF (1995) A Biologically Consistent Model for Comparing Molecular Phylogenies. J Comput Biol 2: 493–507. Available: http://online.liebertpub.com/doi/abs/10.1089/cmb.1995.2.493. Accessed 27 March 2014.

16. Zmasek C, Eddy S (2002) RIO: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. BMC Bioinformatics 3: 14. Available: http://www.biomedcentral.com/1471-2105/3/14. Accessed 27 March 2014.

17. Dehal PS, Boore JL (2006) A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. BMC Bioinformatics 7: 201. Available: http://www.biomedcentral.com/1471-2105/7/201. Accessed 25 March 2014.

18. Altschul SF (1991) Amino acid substitution matrices from an information theoretic perspective. J Mol Biol 219: 555–565. Available: http://www.sciencedirect.com/science/article/pii/002228369190193A. Accessed 8 April 2014.

19. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci 89: 10915–10919. Available: http://www.pnas.org/content/89/22/10915.short. Accessed 8 April 2014.

20. Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. Atlas Protein Seq Struct vol. 5 sup: 345–352. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.145.4315. Accessed 8 April 2014.

21. Jothi R, Zotenko E, Tasneem A, Przytycka TM (2006) COCO-CL: hierarchical clustering of homology relations based on evolutionary correlations. Bioinformatics 22: 779–788. Available: http://bioinformatics.oxfordjournals.org/content/22/7/779.short. Accessed 27 March 2014.

22. Zmasek CM, Eddy SR (2001) A simple algorithm to infer gene duplication and speciation events on a gene tree. Bioinformatics 17: 821–828. Available: http://bioinformatics.oxfordjournals.org/content/17/9/821.abstract. Accessed 27 March 2014.

23. Sennblad B, Lagergren J (2009) Probabilistic orthology analysis. Syst Biol 58: 411–424. Available: http://sysbio.oxfordjournals.org/content/58/4/411.short. Accessed 27 March 2014.

24. Arvestad L, Berglund A-C, Lagergren J, Sennblad B (2003) Bayesian gene/species tree reconciliation and orthology analysis using MCMC. Bioinformatics 19: i7–i15. Available: http://bioinformatics.oxfordjournals.org/content/19/suppl_1/i7.short. Accessed 27 March 2014.

25. Li H, Coghlan A, Ruan J, Coin LJ, Hériché J-K, et al. (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res 34: D572–80. Available: http://nar.oxfordjournals.org/content/34/suppl_1/D572.short. Accessed 24 March 2014.

26. Waterhouse RM, Zdobnov EM, Tegenfeldt F, Li J, Kriventseva E V (2011) OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. Nucleic Acids Res 39: D283–8. Available: http://nar.oxfordjournals.org/content/39/suppl_1/D283.short. Accessed 23 February 2014.

27. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, et al. (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 19: 327–335. Available: http://genome.cshlp.org/content/19/2/327.full. Accessed 27 March 2014.

28. Thornton JW, DeSalle R (2000) Gene family evolution and homology: genomics meets phylogenetics. Annu Rev Genomics Hum Genet 1: 41–73. Available: http://www.annualreviews.org/doi/abs/10.1146/annurev.genom.1.1.41. Accessed 27 March 2014.

29. Koonin E V, Mushegian AR, Bork P (1996) Non-orthologous gene displacement. Trends Genet 12: 334–336. Available: http://www.ncbi.nlm.nih.gov/pubmed/8855656. Accessed 27 March 2014.

30. Thornton JW, Desalle R (2000) A New Method to Localize and Test the Significance of Incongruence: Detecting Domain Shuffling in the Nuclear Receptor Superfamily. Syst Biol 49: 183–201. Available: http://sysbio.oxfordjournals.org/content/49/2/183.short. Accessed 27 March 2014.

31. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29. Available: http://www.nature.com.proxy2.library.illinois.edu/ng/journal/v25/n1/full/ng0500_25.html. Accessed 19 March 2014.

32. Nehrt NL, Clark WT, Radivojac P, Hahn MW (2011) Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Comput Biol 7: e1002073. Available: http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1002073#pcbi-1002073-g004. Accessed 25 March 2014.

33. Wetlaufer DB (1973) Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins. Proc Natl Acad Sci 70: 697–701. Available: http://www.pnas.org/content/70/3/697.short. Accessed 15 April 2014.

34. Bork P (1991) Shuffled domains in extracellular proteins. FEBS Lett 286: 47–54. Available: http://www.sciencedirect.com/science/article/pii/001457939180937X. Accessed 15 April 2014.

35. Letunic I, Doerks T, Bork P (2012) SMART 7: recent updates to the protein domain annotation resource. Nucleic Acids Res 40: D302–5. Available: http://nar.oxfordjournals.org/content/40/D1/D302.short. Accessed 19 March 2014.

36. Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, et al. (2013) New and continuing developments at PROSITE. Nucleic Acids Res 41: D344–7. Available: http://nar.oxfordjournals.org/content/41/D1/D344.short. Accessed 31 March 2014.

37. Lees J, Yeats C, Perkins J, Sillitoe I, Rentzsch R, et al. (2012) Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis. Nucleic Acids Res 40: D465–71. Available: http://nar.oxfordjournals.org/content/40/D1/D465.short. Accessed 9 April 2014.

38. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305: 567–580. Available: http://www.sciencedirect.com/science/article/pii/S0022283600943158. Accessed 19 March 2014.

39. De Lima Morais DA, Fang H, Rackham OJL, Wilson D, Pethica R, et al. (2011) SUPERFAMILY 1.75 including a domain-centric gene ontology method. Nucleic Acids Res 39: D427–34. Available: http://nar.oxfordjournals.org/content/39/suppl_1/D427.short. Accessed 9 April 2014.

40. Jones P, Binns D, Chang H-Y, Fraser M, Li W, et al. (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics: btu031–. Available: http://bioinformatics.oxfordjournals.org/content/early/2014/01/29/bioinformatics.btu031.full. Accessed 19 March 2014.

41. Polymeropoulos MH (1997) Mutation in the α-Synuclein Gene Identified in Families with Parkinson's Disease. Science (80-) 276: 2045–2047. Available: http://www.sciencemag.org/content/276/5321/2045.abstract. Accessed 26 March 2014.

42. Shoichet BK, Baase WA, Kuroki R, Matthews BW (1995) A relationship between protein stability and protein function. Proc Natl Acad Sci 92: 452–456. Available: http://www.pnas.org/content/92/2/452.short. Accessed 8 April 2014.

43. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4: 1073–1081. Available: http://dx.doi.org/10.1038/nprot.2009.86. Accessed 20 March 2014.

44. Yu FH, Yarov-Yarovoy V, Gutman GA, Catterall WA (2005) Overview of molecular relationships in the voltage-gated ion channel superfamily. Pharmacol Rev 57: 387–395. Available: http://pharmrev.aspetjournals.org/content/57/4/387.short. Accessed 14 April 2014.

45. Warmke JW, Ganetzky B (1994) A family of potassium channel genes related to eag in Drosophila and mammals. Proc Natl Acad Sci 91: 3438–3442. Available: http://www.pnas.org/content/91/8/3438.short. Accessed 10 April 2014.

46. Todd AE, Orengo CA, Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 307: 1113–1143. Available: http://www.sciencedirect.com/science/article/pii/S0022283601945139. Accessed 19 March 2014.

47. Kriventseva E V, Koch I, Apweiler R, Vingron M, Bork P, et al. (2003) Increase of functional diversity by alternative splicing. Trends Genet 19: 124–128. Available: http://www.sciencedirect.com/science/article/pii/S0168952503000234. Accessed 28 March 2014.

48. Ishitani A, Geraghty DE (1992) Alternative splicing of HLA-G transcripts yields proteins with primary structures resembling both class I and class II antigens. Proc Natl Acad Sci 89: 3947–3951. Available: http://www.pnas.org/content/89/9/3947.short. Accessed 28 March 2014.

49. Rendon G, Ger M-F, Kantorovitz R, Natarajan S, Tilson J, et al. (2010) Understanding the "Horizontal Dimension" of Molecular Evolution to Annotate, Classify, and Discover Proteins with Functional Domains. J Comput Sci Technol 25: 82–94. Available: http://link.springer.com/10.1007/s11390-010-9307-3. Accessed 12 March 2014.

50. Alaux C, Sinha S, Hasadsri L, Hunt GJ, Guzmán-Novoa E, et al. (2009) Honey bee aggression supports a link between gene regulation and behavioral evolution. Proc Natl Acad Sci 106: 15400–15405.

51. Kantorovitz MR, Robinson GE, Sinha S (2007) A statistical method for alignment-free comparison of regulatory sequences. Bioinformatics 23: i249–i255.

52. Collins T, Stone JR, Williams AJ (2001) All in the family: the BTB/POZ, KRAB, and SCAN domains. Mol Cell Biol 21: 3609–3615. Available: http://mcb.asm.org/content/21/11/3609.short. Accessed 28 March 2014.

53. Stubbs L, Sun Y, Caetano-Anolles D (2011) Function and Evolution of C2H2 Zinc Finger Arrays. Subcell Biochem 52: 75–94. Available: http://www.ncbi.nlm.nih.gov/pubmed/21557079. Accessed 28 March 2014.

54. Pavletich NP, Pabo CO (1991) Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science 252: 809–817. Available: http://www.ncbi.nlm.nih.gov/pubmed/2028256. Accessed 21 February 2014.

55. Kim CA, Berg JM (1996) A 2.2 Å resolution crystal structure of a designed zinc finger protein bound to DNA. Nat Struct Biol 3: 940–945. Available: http://dx.doi.org/10.1038/nsb1196-940. Accessed 21 February 2014.

56. Isalan M, Choo Y, Klug A (1997) Synergy between adjacent zinc fingers in sequence-specific DNA recognition. Proc Natl Acad Sci 94: 5617–5621. Available: http://www.pnas.org/content/94/11/5617.short. Accessed 28 March 2014.

57. Wolfe SA, Nekludova L, Pabo CO (2000) DNA recognition by Cys2His2 zinc finger proteins. Annu Rev Biophys Biomol Struct 29: 183–212. Available: http://www.annualreviews.org/doi/abs/10.1146/annurev.biophys.29.1.183. Accessed 19 March 2014.

58. Oliver PL, Goodstadt L, Bayes JJ, Birtle Z, Roach KC, et al. (2009) Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. PLoS Genet 5: e1000753. Available: http://www.plosgenetics.org/article/info:doi/10.1371/journal.pgen.1000753#pgen-1000753-g007. Accessed 26 March 2014.

59. Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, et al. (2010) PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science 327: 836–840. Available: http://www.sciencemag.org/content/327/5967/836.short. Accessed 20 March 2014.

60. Berg IL, Neumann R, Lam K-WG, Sarbajna S, Odenthal-Hesse L, et al. (2010) PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nat Genet 42: 859–863. Available: http://dx.doi.org/10.1038/ng.658. Accessed 27 March 2014.

61. Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, et al. (2010) Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science 327: 876–879. Available: http://www.sciencemag.org/content/327/5967/876.short. Accessed 20 March 2014.

62. Parvanov ED, Petkov PM, Paigen K (2010) Prdm9 controls activation of mammalian recombination hotspots. Science 327: 835. Available: http://www.sciencemag.org/content/327/5967/835.short. Accessed 28 March 2014.

63. Birtle Z, Ponting CP (2006) Meisetz and the birth of the KRAB motif. Bioinformatics 22: 2841–2845. Available: http://bioinformatics.oxfordjournals.org/content/22/23/2841.short. Accessed 28 March 2014.

64. Consiantinou-Deltas CD, Gilbert J, Bartlett RJ, Herbstreith M, Roses AD, et al. (1992) The identification and characterization of KRAB-domain-containing zinc finger proteins. Genomics 12: 581–589. Available: http://www.sciencedirect.com/science/article/pii/088875439290451W. Accessed 28 March 2014.

65. Bellefroid EJ, Marine JC, Ried T, Lecocq PJ, Rivière M, et al. (1993) Clustered organization of homologous KRAB zinc-finger genes with enhanced expression in human T lymphoid cells. EMBO J 12: 1363–1374. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=413348&tool=pmcentrez&rendertype=abstract. Accessed 28 March 2014.

66. Huntley S, Baggott DM, Hamilton AT, Tran-Gyamfi M, Yang S, et al. (2006) A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. Genome Res 16: 669–677.

67. Margolin JF, Friedman JR, Meyer WK, Vissing H, Thiesen HJ, et al. (1994) Kruppel-associated boxes are potent transcriptional repression domains. Proc Natl Acad Sci 91: 4509–4513. Available: http://www.pnas.org/content/91/10/4509.short. Accessed 20 February 2014.

68. Pengue G, Calabró V, Bartoli PC, Pagliuca A, Lania L (1994) Repression of transcriptional activity at a distance by the evolutionary conserved KRAB domain present in a subfamily of zinc finger proteins. Nucleic Acids Res 22: 2908–2914. Available: http://nar.oxfordjournals.org/content/22/15/2908.short. Accessed 28 March 2014.

69. Witzgall R, O'Leary E, Leaf A, Onaldi D, Bonventre J V. (1994) The Kruppel-associated box-A (KRAB-A) domain of zinc finger proteins mediates transcriptional repression. Proc Natl Acad Sci 91: 4514–4518. Available: http://www.pnas.org/content/91/10/4514.short. Accessed 28 March 2014.

70. Vissing H, Meyer WK-H, Aagaard L, Tommerup N, Thiesen H-J (1995) Repression of transcriptional activity by heterologous KRAB domains present in zinc finger proteins. FEBS Lett 369: 153–157. Available: http://www.sciencedirect.com/science/article/pii/001457939500728R. Accessed 28 March 2014.

71. Nowick K, Fields C, Gernat T, Caetano-Anolles D, Kholina N, et al. (2011) Gain, loss and divergence in primate zinc-finger genes: a rich resource for evolution of gene regulatory differences between species. PLoS One 6: e21553. Available: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0021553#pone-0021553-g004. Accessed 20 February 2014.

72. Okumura K (1997) HUB1, a novel Kruppel type zinc finger protein, represses the human T cell leukemia virus type I long terminal repeat-mediated expression. Nucleic Acids Res 25: 5025–5032. Available: http://nar.oxfordjournals.org/content/25/24/5025.short. Accessed 20 February 2014.

73. Conroy AT, Sharma M, Holtz AE, Wu C, Sun Z, et al. (2002) A novel zinc finger transcription factor with two isoforms that are differentially repressed by estrogen receptor-alpha. J Biol Chem 277: 9326–9334. Available: http://www.jbc.org/content/277/11/9326.long. Accessed 28 March 2014.

74. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947–2948.

75. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The Human Genome Browser at UCSC. Genome Res 12: 996–1006. Available: http://genome.cshlp.org/content/12/6/996.abstract. Accessed 19 February 2014.

76. Looman C, Åbrink M, Mark C, Hellman L (2002) KRAB zinc finger proteins: an analysis of the molecular mechanisms governing their increase in numbers and complexity during evolution. Mol Biol Evol 19: 2118–2130.

77. Elrod-Erickson M, Benson TE, Pabo CO (1998) High-resolution structures of variant Zif268–DNA complexes: implications for understanding zinc finger–DNA recognition. Structure 6: 451–464. Available: http://www.sciencedirect.com/science/article/pii/S0969212698000471. Accessed 21 February 2014.

78. Sokal RR (1958) A statistical method for evaluating systematic relationships. Univ Kans Sci Bull 38: 1409–1438.

79. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59: 307–321. Available: http://sysbio.oxfordjournals.org/content/59/3/307.short. Accessed 21 January 2014.

80. Huerta-Cepas J, Dopazo J, Gabaldón T (2010) ETE: a python Environment for Tree Exploration. BMC Bioinformatics 11: 24.

81. Crooks GE, Hon G, Chandonia J-M, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14: 1188–1190. Available: http://genome.cshlp.org/content/14/6/1188.short. Accessed 20 February 2014.

82. Kim J, Zhao K, Jiang P, Lu Z, Wang J, et al. (2012) Transcriptome landscape of the human placenta. BMC Genomics 13: 115. Available: http://www.biomedcentral.com/1471-2164/13/115. Accessed 28 March 2014.

83. De Hoon MJL, Imoto S, Nolan J, Miyano S (2004) Open source clustering software. Bioinformatics 20: 1453–1454. Available: http://bioinformatics.oxfordjournals.org/content/20/9/1453.short. Accessed 21 March 2014.

84. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 25: 402–408. Available: http://www.sciencedirect.com/science/article/pii/S1046202301912629. Accessed 19 March 2014.

85. Veazey KJ, Golding MC (2011) Selection of stable reference genes for quantitative rt-PCR comparisons of mouse embryonic and extra-embryonic stem cells. PLoS One 6: e27592. Available: http://dx.plos.org/10.1371/journal.pone.0027592. Accessed 28 March 2014.

86. Elso C, Lu X, Weisner PA, Thompson HL, Skinner A, et al. (2013) A reciprocal translocation dissects roles of Pax6 alternative promoters and upstream regulatory elements in the development of pancreas, brain, and eye. Genesis 51: 630–646. Available: http://www.ncbi.nlm.nih.gov/pubmed/23798316. Accessed 28 March 2014.

87. Thomas JH, Schneider S (2011) Coevolution of retroelements and tandem zinc finger genes. Genome Res 21: 1800–1812. Available: http://genome.cshlp.org/content/21/11/1800.short. Accessed 19 March 2014.

88. Corsinotti A, Kapopoulou A, Gubelmann C, Imbeault M, Santoni de Sio FR, et al. (2013) Global and stage specific patterns of Krüppel-associated-box zinc finger protein gene expression in murine early embryonic cells. PLoS One 8: e56721. Available: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0056721#pone-0056721-g005. Accessed 28 March 2014.

89. Dehal P, Predki P, Olsen AS, Kobayashi A, Folta P, et al. (2001) Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. Science 293: 104–111. Available: http://www.sciencemag.org/content/293/5527/104.abstract. Accessed 20 February 2014.

90. Urrutia R (2003) KRAB-containing zinc-finger repressor proteins. Genome Biol 4: 231. Available: http://genomebiology.com/2003/4/10/231. Accessed 20 February 2014.

91. Emerson RO, Thomas JH (2011) Gypsy and the birth of the SCAN domain. J Virol 85: 12043–12052. Available: http://jvi.asm.org/content/85/22/12043.short. Accessed 4 February 2014.

92. Talkowski ME, Rosenfeld JA, Blumenthal I, Pillalamarri V, Chiang C, et al. (2012) Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. Cell 149: 525–537. Available: http://www.sciencedirect.com/science/article/pii/S0092867412004114. Accessed 25 January 2014.

93. Hamilton A, Huntley S, Kim J, Branscomb E, Stubbs L (2003) Lineage-specific Expansion of KRAB Zinc-finger Transcription Factor Genes: Implications for the Evolution of Vertebrate Regulatory Networks. Cold Spring Harb Symp Quant Biol 68: 131–140. Available: http://symposium.cshlp.org/content/68/131.short. Accessed 28 March 2014.

94. Shannon M, Hamilton AT, Gordon L, Branscomb E, Stubbs L (2003) Differential expansion of zinc-finger transcription factor loci in homologous human and mouse gene clusters. Genome Res 13: 1097–1110. Available: http://genome.cshlp.org/content/13/6a/1097.short. Accessed 28 March 2014.

95. Krebs CJ, Larkins LK, Khan SM, Robins DM (2005) Expansion and diversification of KRAB zinc-finger genes within a cluster including Regulator of sex-limitation 1 and 2. Genomics 85: 752–761. Available: http://www.sciencedirect.com/science/article/pii/S0888754305000650. Accessed 28 March 2014.

96. Nowick K, Hamilton AT, Zhang H, Stubbs L (2010) Rapid sequence and expression divergence suggest selection for novel function in primate-specific KRAB-ZNF genes. Mol Biol Evol 27: 2606–2617. Available: http://mbe.oxfordjournals.org/content/27/11/2606.short. Accessed 28 March 2014.

97. Thomas JH, Emerson RO, Shendure J (2009) Extraordinary molecular evolution in the PRDM9 fertility gene. PLoS One 4: e8505. Available: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0008505#pone-0008505-g004. Accessed 27 March 2014.

98. Berg JM (1992) Sp1 and the subfamily of zinc finger proteins with guanine-rich binding sites. Proc Natl Acad Sci U S A 89: 11109–11110. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=50497&tool=pmcentrez&rendertype=abstract. Accessed 28 March 2014.

99. Swamynathan SK (2010) Krüppel-like factors: Three fingers in control. Hum Genomics 4: 263. Available: /pmc/articles/mid/NIHMS244525/?report=abstract. Accessed 28 March 2014.

100. Ding G, Lorenz P, Kreutzer M, Li Y, Thiesen H-J (2009) SysZNF: the C2H2 zinc finger gene database. Nucleic Acids Res 37: D267–73. Available: http://nar.oxfordjournals.org/content/37/suppl_1/D267.short. Accessed 25 March 2014.

101. Emerson RO, Thomas JH (2009) Adaptive evolution in zinc finger transcription factors. PLoS Genet 5: e1000325. Available: http://www.plosgenetics.org/article/info:doi/10.1371/journal.pgen.1000325#pgen-1000325-g006. Accessed 28 March 2014.

102. Quenneville S, Turelli P, Bojkowska K, Raclot C, Offner S, et al. (2012) The KRAB-ZFP/KAP1 system contributes to the early embryonic establishment of site-specific DNA methylation patterns maintained during development. Cell Rep 2: 766–773. Available: http://www.sciencedirect.com/science/article/pii/S2211124712003166. Accessed 28 March 2014.

103. Santoni de Sio FR, Barde I, Offner S, Kapopoulou A, Corsinotti A, et al. (2012) KAP1 regulates gene networks controlling T-cell development and responsiveness. FASEB J 26: 4561–4575. Available: http://www.fasebj.org/content/26/11/4561.short. Accessed 28 March 2014.

104. Santoni de Sio FR, Massacand J, Barde I, Offner S, Corsinotti A, et al. (2012) KAP1 regulates gene networks controlling mouse B-lymphoid cell differentiation and function. Blood 119: 4675–4685. Available: http://bloodjournal.hematologylibrary.org/content/119/20/4675.short. Accessed 28 March 2014.

105. Barde I, Rauwel B, Marin-Florez RM, Corsinotti A, Laurenti E, et al. (2013) A KRAB/KAP1-miRNA cascade regulates erythropoiesis through stage-specific control of mitophagy. Science 340: 350–353. Available: http://www.sciencemag.org/content/340/6130/350.short. Accessed 28 March 2014.

106. Schep AN, Adryan B (2013) A comparative analysis of transcription factor expression during metazoan embryonic development. PLoS One 8: e66826. Available: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0066826#pone-0066826-g006. Accessed 27 March 2014.

107. Hui C-C, Angers S (2011) Gli proteins in development and disease. Annu Rev Cell Dev Biol 27: 513–537. Available: http://www.annualreviews.org/doi/abs/10.1146/annurev-cellbio-092910-154048. Accessed 27 March 2014.

108. Ali RG, Bellchambers HM, Arkell RM (2012) Zinc fingers of the cerebellum (Zic): transcription factors and co-factors. Int J Biochem Cell Biol 44: 2065–2068. Available: http://www.sciencedirect.com/science/article/pii/S1357272512002907. Accessed 28 March 2014.

109. Siggs OM, Beutler B (2012) The BTB-ZF transcription factors. Cell Cycle 11: 3358–3369. Available: https://www.landesbioscience.com/journals/cc/article/21277/. Accessed 28 March 2014.

110. Zhao C, Meng A (2005) Sp1-like transcription factors are regulators of embryonic development in vertebrates. Dev Growth Differ 47: 201–211. Available: http://www.ncbi.nlm.nih.gov/pubmed/15921495. Accessed 28 March 2014.

111. Black SG, Arnaud F, Palmarini M, Spencer TE (2010) Endogenous retroviruses in trophoblast differentiation and placental development. Am J Reprod Immunol 64: 255–264. Available: http://www.ncbi.nlm.nih.gov/pubmed/20528833. Accessed 28 March 2014.

112. Lindenfors P, Tullberg BS (2011) Evolutionary aspects of aggression the importance of sexual selection. Adv Genet 75: 7–22. Available: http://www.sciencedirect.com/science/article/pii/B9780123808585000095. Accessed 28 March 2014.

113. Moore T (2012) Review: Parent-offspring conflict and the control of placental function. Placenta 33 Suppl: S33–6. Available: http://www.sciencedirect.com/science/article/pii/S0143400411005480. Accessed 28 March 2014.

114. McPherson FJ, Chenoweth PJ (2012) Mammalian sexual dimorphism. Anim Reprod Sci 131: 109–122. Available: http://www.sciencedirect.com/science/article/pii/S0378432012000619. Accessed 27 March 2014.

115. Robinson GE, Fernald RD, Clayton DF (2008) Genes and social behavior. Science 322: 896–900. Available: http://www.sciencemag.org/content/322/5903/896.short. Accessed 25 March 2014.

116. Newman SW (1999) The Medial Extended Amygdala in Male Reproductive Behavior A Node in the Mammalian Social Behavior Network. Ann N Y Acad Sci 877: 242–257. Available: http://doi.wiley.com/10.1111/j.1749-6632.1999.tb09271.x. Accessed 31 March 2014.

117. Mello C V., Vicario DS, Clayton DF (1992) Song presentation induces gene expression in the songbird forebrain. Proc Natl Acad Sci 89: 6818–6822. Available: http://www.pnas.org/content/89/15/6818.short. Accessed 31 March 2014.

118. Burmeister SS, Jarvis ED, Fernald RD (2005) Social Opportunity Produces Brain Changes in Fish. PLoS Biol 3: e390. Available: http://dx.plos.org/10.1371/journal.pbio.0030390. Accessed 31 March 2014.

119. Weaver ICG, Cervoni N, Champagne FA, D'Alessio AC, Sharma S, et al. (2004) Epigenetic programming by maternal behavior. Nat Neurosci 7: 847–854. Available: http://dx.doi.org/10.1038/nn1276. Accessed 19 March 2014.

120. Maruska KP, Fernald RD (2011) Social regulation of gene expression in the hypothalamic-pituitary-gonadal axis. Physiology (Bethesda) 26: 412–423. Available: http://physiologyonline.physiology.org/content/26/6/412.short. Accessed 31 March 2014.

121. Robinson GE, Grozinger CM, Whitfield CW (2005) Sociogenomics: social life in molecular terms. Nat Rev Genet 6: 257–270.

122. Mikkelsen TS, Hillier LW, Eichler EE, Zody MC, Jaffe DB, et al. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature.

123. Prüfer K, Munch K, Hellmann I, Akagi K, Miller JR, et al. (2012) The bonobo genome compared with the chimpanzee and human genomes. Nature.

124. Diamond JM (2006) The third chimpanzee: The evolution and future of the human animal. HarperCollins.

125. Weinstock GM, Robinson GE, Gibbs RA, Worley KC, Evans JD, et al. (2006) Insights into social insects from the genome of the honeybee Apis mellifera. Nature 443: 931–949.

126. Grimaldi D (2005) Evolution of the Insects. Cambridge University Press.

127. Fang G, Bhardwaj N, Robilotto R, Gerstein MB (2010) Getting started in gene orthology and functional analysis. PLoS Comput Biol 6: e1000703.

128. Hunt GJ (2007) Flight and fight: a comparative view of the neurophysiology and genetics of honey bee defensive behavior. J Insect Physiol 53: 399–410.

129. Ben-Shahar Y, Robichon A, Sokolowski MB, Robinson GE (2002) Influence of gene action across different time scales on behavior. Science (80-) 296: 741–744.

130. Fisher SRA, Genetiker S, Fisher RA, Genetician S, Britain G, et al. (1970) Statistical methods for research workers. Oliver and Boyd Edinburgh.

131. Da Wei Huang BTS, Lempicki RA, others, Huang DW, Sherman BT (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44–57. Available: http://www.nature.com.proxy2.library.illinois.edu/nprot/journal/v4/n1/full/nprot.2008.211.html. Accessed 19 March 2014.

132. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33: D514–D517.

133. Wu CH, Huang H, Nikolskaya A, Hu Z, Barker WC (2004) The iProClass integrated database for protein functional analysis. Comput Biol Chem 28: 87–96.

134. Bairoch A, Bougueleret L, Altairac S, Amendolia V, Auchincloss A, et al. (2008) The universal protein resource (UniProt). Nucleic Acids Res 36: D190–D195.

135. Becker KG, White SL, Muller J, Engel J (2000) BBID: the biological biochemical image database. Bioinformatics 16: 745–746.

136. Nishimura D (2001) BioCarta. Biotech Softw Internet Rep Comput Softw J Sci 2: 117–120.

137. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30.

138. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, et al. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res 29: 37–40.

139. Wu CH, Nikolskaya A, Huang H, Yeh L-SL, Natale DA, et al. (2004) PIRSF: family classification system at the Protein Information Resource. Nucleic Acids Res 32: D112–4. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=308831&tool=pmcentrez&rendertype=abstract. Accessed 19 September 2013.

140. Schultz J, Copley RR, Doerks T, Ponting CP, Bork P (2000) SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res 28: 231–234.

141. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498–2504. Available: http://genome.cshlp.org/content/13/11/2498.short. Accessed 13 December 2013.

142. Kantorovitz MR, Kazemian M, Kinston S, Miranda-Saavedra D, Zhu Q, et al. (2009) Motif-Blind, Genome-Wide Discovery of cis-Regulatory Modules in Drosophila and Mouse. Dev Cell 17: 568–579. Available: http://www.sciencedirect.com/science/article/pii/S1534580709003840. Accessed 30 September 2013.

143. Kantorovitz MR, Tcheng D, Lerman MI, Jakobsson E (2013) Genome wide identification of regulatory networks associated with general cognitive ability using a normalized alignment free similarity measure of promoter regions. arXiv Prepr arXiv13041231.

144. Wakaguri H, Yamashita R, Suzuki Y, Sugano S, Nakai K (2008) DBTSS: database of transcription start sites, progress report 2008. Nucleic Acids Res 36: D97–D101.

145. Mega T, Lupia M, Amodio N, Horton SJ, Mesuraca M, et al. (2011) Zinc finger protein 521 antagonizes early B-cell factor 1 and modulates the B-lymphoid differentiation of primary hematopoietic progenitors. Cell Cycle 10: 2129–2139.

146. Jian X-Q, Wang K-S, Wu T-J, Hillhouse JJ, Mullersman JE (2011) Association of ADAM10 and CAMK2A polymorphisms with conduct disorder: Evidence from family-based studies. J Abnorm Child Psychol 39: 773–782.

147. Shen S, Pu J, Lang B, McCaig CD, others (2011) A zinc finger protein Zfp521 directs neural differentiation and beyond. Stem Cell Res Ther 2: 1–4.

148. Noble EP (2003) D2 dopamine receptor gene in psychiatric and neurologic disorders and its phenotypes. Am J Med Genet Part B Neuropsychiatr Genet 116: 103–125.

149. Ishikawa-Brush Y, Powell JF, Bolton P, Miller AP, Francis F, et al. (1997) Autism and multiple exostoses associated with an X; 8 translocation occurring within the GRPR gene and 3’ to the SDC2 gene. Hum Mol Genet 6: 1241–1250.

150. Innos J, Philips M-A, Raud S, Lilleväli K, Kõks S, et al. (2012) Deletion of the Lsamp gene lowers sensitivity to stressful environmental manipulations in mice. Behav Brain Res 228: 74–81. Available: http://www.sciencedirect.com/science/article/pii/S0166432811008321. Accessed 30 September 2013.

151. Must A, Tasa G, Lang A, Vasar E, Kõks S, et al. (2008) Association of limbic system-associated membrane protein (LSAMP) to male completed suicide. BMC Med Genet 9: 34.

152. Narayan S, Head SR, Gilmartin TJ, Dean B, Thomas EA (2009) Evidence for disruption of sphingolipid metabolism in schizophrenia. J Neurosci Res 87: 278–288.

153. Cuscó I, Medrano A, Gener B, Vilardell M, Gallastegui F, et al. (2009) Autism-specific copy number variants further implicate the phosphatidylinositol signaling pathway and the glutamatergic synapse in the etiology of the disorder. Hum Mol Genet 18: 1795–1804.

154. Matsumoto M, Straub RE, Marenco S, Nicodemus KK, Matsumoto S, et al. (2008) The evolutionarily conserved G protein-coupled receptor SREB2/GPR85 influences brain size, behavior, and vulnerability to schizophrenia. Proc Natl Acad Sci 105: 6133–6138.

155. Tasneem A, Iyer LM, Jakobsson E, Aravind L (2005) Identification of the prokaryotic ligand-gated ion channels and their implications for the mechanisms and origins of animal Cys-loop ion channels. Genome Biol 6: R4. Available: http://genomebiology.com/2004/6/1/R4. Accessed 9 April 2014.

156. Rendon G, Kantorovitz MR, Tilson JL, Jakobsson E (2011) Identifying bacterial and archaeal homologs of pentameric ligand-gated ion channel (pLGIC) family using domain-based and alignment-based approaches. Channels (Austin) 5: 325–343. Available: https://www.landesbioscience.com/journals/channels/article/16822/?nocache=192075600. Accessed 12 March 2014.

157. Bailey TL, Elkan C, others (1994) Fitting a mixture model by expectation maximization to discover motifs in bipolymers.

158. Bailey TL, Gribskov M (1998) Combining evidence using p-values: application to sequence homology searches. Bioinformatics 14: 48–54. Available: http://bioinformatics.oxfordjournals.org/content/14/1/48.short. Accessed 14 April 2014.

159. Tilson JL, Rendon G, Ger M-F, Jakobsson E (2007) MotifNetwork: A grid-enabled workflow for high-throughput domain analysis of biological sequences: Implications for annotation and study of phylogeny, protein interactions, and intraspecies variation. Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on. pp. 620–627.

160. WANG L, JIANG T (1994) On the Complexity of Multiple Sequence Alignment. J Comput Biol 1: 337–348. Available: http://online.liebertpub.com/doi/abs/10.1089/cmb.1994.1.337. Accessed 11 April 2014.

161. Elias DI (2006) Settling the Intractability of Multiple Alignment. Available: http://online.liebertpub.com/doi/abs/10.1089/cmb.2006.13.1323. Accessed 11 April 2014.

162. Elsik CG, Worley KC, Bennett AK, Beye M, Camara F, et al. (2014) Finding the missing honey bee genes: lessons learned from a genome upgrade. BMC Genomics 15: 86. Available: http://www.biomedcentral.com/1471-2164/15/86. Accessed 27 February 2014.

163. Chandrasekaran S, Ament SA, Eddy JA, Rodriguez-Zas SL, Schatz BR, et al. (2011) Behavior-specific changes in transcriptional modules lead to distinct and predictable neurogenomic states. Proc Natl Acad Sci U S A 108: 18020–18025. Available: http://www.pnas.org/content/108/44/18020.short. Accessed 6 February 2014.

164. Liu H, Chang L-H, Sun Y, Lu X, Stubbs L (2014) Deep vertebrate roots for Mammalian zinc finger transcription factor subfamilies. Genome Biol Evol 6: 510–525. Available: http://gbe.oxfordjournals.org/content/6/3/510.short. Accessed 28 March 2014.

165. Kiontke K, Fitch DHA (2005) The phylogenetic relationships of Caenorhabditis and other rhabditids.

166. Miller W, Rosenbloom K, Hardison RC, Hou M, Taylor J, et al. (2007) 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. Genome Res 17: 1797–1808.

Tables and Figures

Tables

Table 1. ZNF gene models in each subfamily with and without Ensembl model overlap

	ZNF-only		BTB/POZ		KRAB		SCAN		SCAN-KRAB	
Species	**All**	**Ensembl**	**All**	**Ensembl**	**All**	**Ensembl**	**All**	**Ensembl**	**All**	**Ensembl**
Human [1]	212	212	42	42	366	366	29	29	28	28
Mouse	590	353	40	39	523	370	27	23	14	14
Opossum	868	548	23	23	709	513	0	0	20	19
Chicken	290	219	23	23	47	39	0	0	0	0
Zebra finch	1026	749	23	22	3	2	0	0	0	0
Lizard	723	484	30	28	122	67	89	54	240	126
Frog	473	293	34	33	158	112	0 [2]	0	0	0

[1] Manually curated protein-coding loci from[66] and including human orthologs of mouse or other species genes from Genbank, as described in the text.

[2] Excluding the single model also detected by[91] but considered a false prediction, as discussed in the text.

Table 2. Predicted novel mouse genes with excellent fingerprint match in other species

Model	Conserved in [1]	Human match	Type	Example EST/mRNA (accession number) [2]	Example EST source(s)
ZF01023_1	Md, Hs	GLI4	Zx5	AK084954	Whole embryo; E14.5 haematopoietic
ZF04524_2	Cl, Hs	ZNF471	ABZx14	M36516	Oocyte; embryo eye
ZF02332_2	Cl, Hs	ZNF582	ABZx9	BB619218; BU054342	E8 whole; E12.5 brain
ZF02433_2	Cl, Hs	ZNF570	ABZx9	AK138949; CJ048012	Aorta; 11d pregnant uterus
ZF02869_1	Cl, Hs	ZNF660	Zx10	BB193415; CB524555	Spinal cord; E12.5 brain
ZF04379_1	Gg, Md, Cl, Hs	ZNF853	Zx5	BG277278	Maxillary process
ZF02438_1	Cl, Hs	ZNF567	AZx13	None	no EST
ZF02320_1	Cl, Hs	ZNF331	Zx8	None	no EST
ZF05506_1	Md, Cl, Hs	ZNF16	Zx9	None	no EST

[1] Gg, chicken; Md, opossum; Cl, dog; Hs, human.

[2] Overlapping ESTs; only example ESTs are listed, other ESTs overlap with most models.

Table 3. Predicted novel mouse genes that are not conserved in other species (selected examples)

Example Model	Cluster location	# clustered genes; known [1]	Type	Example EST	EST sources
ZF00123_1	Chr1:119-121 Mb	18 genes;	Zx17	AK139669	2-cell embryo
ZF00134_1		none known	Zx16	DV654250	oocyte
ZF00304_2	Chr10:81 Mb	20 genes; Zfp873 known	ABZx17	CK635639	E9.5-10.5 upper head
ZF00313_2			ABZx16	CF725361	mid-gestation embryo eye
ZF00529_2			ABZx12	BU519096; CJ049410	Undifferentiated limb; E13 testis
ZF00537_2			ABZx14	BB452393	E12 spinal ganglion
ZF00548_2	Chr12:18-25 Mb	51 genes; none known	ABZx18	AV579126	ES cells
ZF00551_2			ABZx15	BQ551390	Mixed adult tissue
ZF06013_2			ABZx13	BM201758; BF714015	E7.5 whole embryo; E10.5 branchial arches
ZF04218_1			Zx17	AU017585; DV645475	2-cell embryo, occyte
ZF04215_1			Zx11	AK136154;	In vitro fertilized egg, oocyte
	Chr6:130.4-	14 genes; none known		DV649857	
ZF04205_1	131.2 Mb		Zx17	BX513671	2-cell embryo
ZF04202_1			Zx13	DV65065	oocyte
ZF04199_1			Zx14	CA559522	Unfertilized egg
ZF04196_1			Zx13	BG080473	Mixed tissue

[1] One example of a Refseq ZNF gene is shown here for each cluster although several cluster members may be known.

Table 4. Different types of human (Homo sapiens) C2H2 genes that are conserved across different domains.

Conserved In	BTB/POZ	KRAB	SCAN	SCANKRAB	ZNF	Sum
Human	0	185	4	1	18	208
Eutherians	0	160	15	10	28	213
Mammals	0	17	10	17	19	63
Amniotes	4	3	0	0	11	18
Tetrapods	38	1	0	0	136	175
Sum	42	366	29	28	212	677

Table 5. Fingerprint alignments of ZNF777, ZNF282 and ZNF507.

ZNF777	Fingerprints								
Human	LINI	HQQL	HSSK	LISI	RSHR	ESKN	RHHE	QHHE	YQSY
Dog	LINI	HQQL	HSSK	LISI	RSHR	ESKN	RHHE	QHHE	YQSY
Mouse	LINI	HQQL	HSSK	LISI	RSHR	ESKN	RHHE	QHHE	-
Opossum	LINI	HQQL	HSSK	LISI	RSHR	ESKN	RHHE	QHHE	-
Finch	LINI	NQQL	HSSK	LISM	RSHR	ESKN	RHHE	QHHE	-
Lizard	LINI	IQQL	HSSK	LISM	RSHR	ESKN	RHHE	QHHE	-
Chick	-	NQQL	HSSK	LISM	RSHR	ESKN	RHHE	QHHE	-

ZNF282	Fingerprints					
Human	VKSI	CSGR	REHN	RQNK	YESD	
Dog	VKSI	CSGR	REHN	RQNK	YESD	
Mouse	VKSI	CSGR	REHN	RQNK	YESD	
Opossum	VKSI	CSGR	REHN	RQNK	YESD	
Lizard	VKSV	CSGR	REHN	RQNK	YESD	

ZNF507	Fingerprints									
Human	SSFL	-	QRMT	NGYQ	NKDS	YSQN	TVGN	-	HPSS	SEND
Mouse	SSFL	-	QRMT	NGYQ	NKDS	STYV	-	-	HPSS	SESD
Oposm	SSLL	-	QRMT	NGYQ	NKDS	-	-	-	-	-
Finch	SSLL	SEES	QRMT	NGYQ	DKGS	-	-	-	-	-
Chick	SSLL	-	QRMT	NGYQ	NKDS	-	-	-	-	-
Lizard	PLPK	-	QRMT	NGYQ	NKDS	-	TVGN	NNSC	HPSS	NEHD
Frog	SETI	KEDG	QRMT	NGYQ	NSDN	STYV	-	NSSC	HPSS	SELE

Table 6. Summary of the sets of differentially expressed genes analyzed in this study [a]

Set Number and Name	Number of genes	Number of genes mapable[1]	Set Description
1. Alarm_Pheromone (large behavioral phenotype difference)	344	275	European bees exposed to alarm pheromone vs European control bees
2. Old_vs_Young (large behavioral phenotype difference)	1125	899	European old bees vs European young bees
3. Soldier_CG (large behavioral phenotype difference)	664	512	African colony soldier bees vs European colony soldier bees
4. Soldier_WG (large behavioral phenotype difference)	396	308	Genetically African soldier bees vs genetically European soldier bees
5. Forager_CG (smaller behavioral phenotype difference)	236	180	African colony forager bees vs European colony forager bees
6. Forager_WG (smaller behavioral phenotype difference)	41	22	Genetically African forager bees vs genetically European forager bees
7. Guard_CG (smaller behavioral phenotype difference)	336	248	African colony guard bees vs European colony guard bees
8. Guard_WG (smaller behavioral phenotype difference)	173	132	Genetically African guard bees vs genetically European guard bees

[a] All eight sets in this table are from[50]. For each social class of bee (forager, guard, soldier) there are four subpopulations: AE (Genetically African bees in European colony), AA (Genetically African bees in African colony), EA (Genetically European bees in African colony), EE (Genetically European bees in European colony). For the sets labeled "WG" (Worker Genotype) AE and AA are integrated via ANOVA statistics into one set and compared to the integrated set comprised of EE and EA. For the sets labeled "CG" (Colony Genotype) AA and EA are integrated via ANOVA into one set and compared to the integrated set comprised of AE and EE. Sets 1, 2, 3, 4 are associated with very large behavioral differences in aggression during hive defense. Sets 5, 6, 7, 8 are sets associated with smaller behavioral differences.

[1] "Number of genes mapable" are the number of differentially expressed genes whose IDs are mapable to InParanoid ortholog database.

Table 7. Statistics of the ortholog count data of sets of differentially expressed honey bee genes [a]

Set Name	*Total Number of Orthologs*	*P value*	*Set Size*	*Average Number of Orthologs per gene*	*Standard Deviation*
Alarm_Pheromone	9402	**0.011539***	275	34.19	16.99
Old_vs_Young	29722	**0.011921***	899	33.06	17.61
Soldier_CG	17062	**0.022775***	512	33.32	16.65
Soldier_WG	9461	0.861547	308	30.72	18.43
Forager_CG	5403	0.911358	180	30.02	17.75
Forager_WG	487	0.993237	22	22.14	18.48
Guard_CG	7098	0.997159	248	28.62	17.63
Guard_WG	3529	0.999360	132	26.73	17.67

[a] Names of sets of differentially expressed genes are the same as tabulated in Table 6. The p-values for the mean number of orthologs are calculated by random sampling, see statistics part of Methods. The three sets selectively enriched in genes conserved across the metazoan are bolded and marked by asterisk.

Table 8. Number of Honey Bee genes that have orthologs in all vertebrate species

Set Name	*Original Set Size*	*Mappable Set Size*[1]	*Conserved*[2]	*P-value*[3]
Alarm_Pheromone	344	275	85	0.0487
Old_vs_Young	1125	899	240	0.486
Soldier_CG	664	512	118	0.972
Soldier_WG	396	308	87	0.244
Forager_CG	236	180	46	0.601
Forager_WG	41	22	2	0.958
Guard_CG	336	248	55	0.944
Guard_WG	173	132	28	0.912

[1] Mappable set size is the size of the set that can be mapped to Inparanoid database.

[2] Conserved column are the number of genes in mappable set that have ortholog in all vertebrate species(altogether 19, ranging from T.nigroviridis to H.sapiens) that included in Inparanoid.

[3] P-values are calculated by hypergeometric test.

Table 9. Exact GO term names for GO Biological Process Tree for the 'Vertebrate-conserved' Alarm_Pheromone Genes' Human Orthologs (Part 1) [a]

Index[1]	*ID*	*Term*[2]	*Pvalue*	*Parent*[3]	*Children*[3]	*Set_Desc*[4]
35	GO:0008150	biological_process	>0.05		118,67,16,116,7,77,11,111	Parent Node
118	GO:0071840	cellular component organization or biogenesis	>0.05	35	58,2	Parent Node
58	GO:0016043	cellular component organization**	0.005362	118	95,23	Up+Combined
23	GO:0022607	cellular component assembly**	6.13E-04	58		Up+Combined
95	GO:0043933	macromolecular complex subunit organization**	1.82E-05	58	26,37	Up+Combined
26	GO:0065003	macromolecular complex assembly**	4.44E-04	95	82	Up+Combined
82	GO:0034622	cellular macromolecular complex assembly**	0.001886	26	79	Up+Combined
79	GO:0043623	cellular protein complex assembly**	0.001174	82	108	Up+Combined
108	GO:0051131	chaperone-mediated protein complex assembly*	0.020886	79		Up+Combined
37	GO:0071822	protein complex subunit organization	>0.05	95	78	Parent Node
78	GO:0006461	protein complex assembly**	4.60E-04	37		Up+Combined
2	GO:0044085	cellular component biogenesis**	0.001378	118	56	Up+Combined
56	GO:0070271	protein complex biogenesis**	4.40E-04	2		Up+Combined
111	GO:0051179	localization*	0.047276	35		Combined
11	GO:0051234	establishment of localization	>0.05	35	18	Parent Node
18	GO:0051649	establishment of localization in cell*	0.040771	11		Combined
116	GO:0065007	biological regulation	>0.05	35	49	Parent Node
49	GO:0050789	regulation of biological process	>0.05	116	57	Parent Node
57	GO:0050794	regulation of cellular process	>0.05	49	34	Parent Node
34	GO:0051726	regulation of cell cycle*	0.013716	57		Up

[a] This table corresponds to the left part of the tree in Figure 12 and is designed to correspond to the topology of the tree. Thus entries in the table follow the vertical lineages in the tree starting with the left-most vertical lineage.

[1] "Index" column is the index number used in Figure 12.

[2] 'Term' column is marked according to p-value significant level: p-values that are $\leq$ 0.01 are marked "**", p-values that are between 0.01 and 0.05 are marked "*".

[3] 'Parent' and 'Children' columns list the indexes of the parent/children node(s) of each node.

[4] 'Set_Desc' column delineates whether the enrichment is among the up-regulated, down-regulated, or up- and down-regulated components combined of the differentially regulated set. Terms of "Parent Node" are terms that are parent node of other enriched GO terms.

Table 10. Exact GO term names for GO Biological Process Tree for the 'Vertebrate-conserved' Alarm_Pheromone Genes' Human Orthologs (Part 2) [a]

Index	*ID*	*Term*	*Pvalue*	*Parent*	*Children*	*Set_Desc*
77	GO:0009987	cellular process*	0.021665	35	9	Combined
9	GO:0044237	cellular metabolic process**	0.008495	77	103,19	Combined
103	GO:0006793	phosphorus metabolic process	>0.05	9	10	Parent Node
10	GO:0006796	phosphate-containing compound metabolic process	>0.05	103	59	Parent Node
59	GO:0016310	phosphorylation*	0.030734	10		Up+Combined
19	GO:0006091	generation of precursor metabolites and energy*	0.015619	9		Down
67	GO:0050896	response to stimulus	>0.05	35	50,87,12,62	Parent Node
62	GO:0009628	response to abiotic stimulus*	0.018963	67		Up
87	GO:0006950	response to stress**	0.001588	67		Up
50	GO:0042221	response to chemical stimulus	>0.05	67	92	Parent Node
92	GO:0010033	response to organic substance*	0.014722	50	17	Up
17	GO:0035966	response to topologically incorrect protein	>0.05	92	83	Parent Node
83	GO:0006986	response to unfolded protein**	4.58E-05	17		Up+Combined
12	GO:0009607	response to biotic stimulus**	0.004111	67		Up+Combined

[a] This table corresponds to the middle part of the tree in Figure 12. Columns are defined as in Table 9.

Table 11. Exact GO term names for GO Biological Process Tree for the 'Vertebrate-conserved' Alarm_Pheromone Genes' Human Orthologs (Part 3) [a]

Index	*ID*	*Term*	*Pvalue*	*Parent*	*Children*	*Set_Desc*
7	GO:0044699	single-organism process	>0.05	35	40	Parent Node
40	GO:0044763	single-organism cellular process	>0.05	7	36,104	Parent Node
36	GO:0051716	cellular response to stimulus	>0.05	40	15	Parent Node
15	GO:0007165	signal transduction	>0.05	36	27	Parent Node
27	GO:0030522	intracellular receptor-mediated signaling pathway**	0.005819	15		Up+Combined
104	GO:0006996	organelle organization*	0.028668	40	74,101	Up+Combined
74	GO:0007010	cytoskeleton organization	>0.05	104	68	Parent Node
68	GO:0030865	cortical cytoskeleton organization*	0.039653	74	38	Up
38	GO:0030866	cortical actin cytoskeleton organization*	0.036872	68		Up
101	GO:0051276	chromosome organization	>0.05	104	53	Parent Node
53	GO:0006325	chromatin organization*	0.030367	101		Combined
16	GO:0008152	metabolic process*	0.036585	35	69,94	Combined
69	GO:0071704	organic substance metabolic process	>0.05	16	28,43	Parent Node
28	GO:0043170	macromolecule metabolic process	>0.05	69	51,30	Parent Node
30	GO:0019538	protein metabolic process*	0.017315	28		Up+Combined
51	GO:0044260	cellular macromolecule metabolic process	>0.05	28	63	Parent Node
63	GO:0044267	cellular protein metabolic process**	0.003561	51	14	Up+Combined
14	GO:0006457	protein folding**	1.76E-04	63		Up+Combined
43	GO:0005975	carbohydrate metabolic process	>0.05	69	47	Parent Node
47	GO:0044723	single-organism carbohydrate metabolic process	>0.05	43	73,84	Parent Node
84	GO:0044724	single-organism carbohydrate catabolic process	>0.05	47	1	Parent Node
1	GO:0046365	monosaccharide catabolic process*	0.042734	84	109	Combined
109	GO:0019320	hexose catabolic process*	0.044613	1	99	Combined
99	GO:0006007	glucose catabolic process	>0.05	109	3	Parent Node
3	GO:0006096	glycolysis*	0.019947	99		Combined
73	GO:0005996	monosaccharide metabolic process	>0.05	47	119	Parent Node
119	GO:0019318	hexose metabolic process*	0.040323	73	13	Down
13	GO:0006006	glucose metabolic process*	0.030211	119		Down+Combined
94	GO:0044238	primary metabolic process*	0.04913	16		Up

[a] This table corresponds to the right part of the tree in Figure 12. Columns are defined as in Table 9.

Table 12. Exact GO term names for GO Molecular Function Tree of the 'Vertebrate-conserved' Alarm_Pheromone Genes' Human Orthologs [a]

Index	*ID*	*Term*	*Pvalue*	*Parent*	*Children*	*Set_Desc*
41	GO:0003674	molecular_function	>0.05		25,21	Parent Node
21	GO:0003824	catalytic activity*	0.035532	41	76,90,48	Combined
90	GO:0016829	lyase activity*	0.032088	21	24	Combined
24	GO:0016830	carbon-carbon lyase activity	>0.05	90	112	Parent Node
112	GO:0016832	aldehyde-lyase activity*	0.02422	24		Combined
76	GO:0016740	transferase activity	>0.05	21	117	Parent Node
117	GO:0016772	transferase activity, transferring phosphorus-containing groups	>0.05	76	4,71	Parent Node
71	GO:0016301	kinase activity*	0.023007	117		Up
4	GO:0016773	phosphotransferase activity, alcohol group as acceptor*	0.011916	117	8	Up+Combined
8	GO:0004672	protein kinase activity**	0.004883	4		Up+Combined
48	GO:0016853	isomerase activity	>0.05	21	75	Parent Node
75	GO:0016866	intramolecular transferase activity	>0.05	48	89	Parent Node
89	GO:0016868	intramolecular transferase activity, phosphotransferases*	0.047865	75		Combined
25	GO:0005488	binding**	0.003461	41	65,70,52	Up+Combined
65	GO:0005515	protein binding**	0.001298	25	55,113,86,61	Up+Combined
113	GO:0019899	enzyme binding	>0.05	65	93	Parent Node
93	GO:0019900	kinase binding*	0.012404	113		Up+Combined
61	GO:0005102	receptor binding	>0.05	65	66	Parent Node
66	GO:0051427	hormone receptor binding*	0.049918	61	105	Combined
105	GO:0035257	nuclear hormone receptor binding*	0.039853	66		Combined
55	GO:0051082	unfolded protein binding**	3.35E-04	65		Up+Combined
86	GO:0031072	heat shock protein binding*	0.019546	65		Up+Combined
52	GO:0060090	binding, bridging*	0.015421	25	85,102	Up+Combined
85	GO:0030674	protein binding, bridging**	0.002543	52		Up+Combined
102	GO:0035591	signaling adaptor activity	>0.05	52	97	Parent Node
97	GO:0005070	SH3/SH2 adaptor activity**	0.007773	102		Up+Combined
70	GO:1901363	heterocyclic compound binding	>0.05	25	60,39	Parent Node
39	GO:0001882	nucleoside binding**	0.001289	70	110	Up+Combined
110	GO:0001883	purine nucleoside binding**	0.001407	39		Up+Combined
60	GO:1901265	nucleoside phosphate binding	>0.05	70	42	Parent Node
42	GO:0000166	nucleotide binding**	0.002572	60	80,22	Up+Combined
80	GO:0017076	purine nucleotide binding**	0.002122	42	31	Up+Combined
31	GO:0030554	adenyl nucleotide binding**	5.40E-04	80		Up+Combined
22	GO:0032553	ribonucleotide binding**	0.004537	42	107	Up+Combined
107	GO:0032555	purine ribonucleotide binding**	0.002132	22	32	Up+Combined
32	GO:0032559	adenyl ribonucleotide binding**	0.001084	107		Up+Combined

[a] 'Index' column is the index number used in the Figure 13. 'Term', 'Parent', 'Children' and 'Set_Desc' columns are defined as in Table 9.

Table 13. Exact GO term names for GO Cellular Component Tree for the 'Vertebrate-conserved' Alarm_Pheromone Genes' Human Orthologs [a]

Index	*ID*	*Term*	*Pvalue*	*Parent*	*Children*	*Set_Desc*
45	GO:0005575	cellular_component	>0.05		72,96,6	Parent Node
96	GO:0043226	organelle**	2.55E-04	45	29	Up+Down+Combined
29	GO:0043227	membrane-bounded organelle**	0.001124	96		Down+Combined
72	GO:0044464	cell part	>0.05	45	64,20	Parent Node
64	GO:0044424	intracellular part**	1.90E-04	72	5,114,81,88	Up+Down+Combined
5	GO:0005737	cytoplasm**	8.11E-05	64		Up+Down+Combined
88	GO:0044446	intracellular organelle part	>0.05	64	115	Parent Node
115	GO:0044428	nuclear part	>0.05	88	33	Parent Node
33	GO:0044451	nucleoplasm part*	0.046018	115		Combined
114	GO:0044444	cytoplasmic part*	0.016664	64	91	Down+Combined
91	GO:0005829	cytosol**	0.001108	114		Up+Combined
81	GO:0043229	intracellular organelle**	4.88E-04	64	46	Up+Down+Combined
46	GO:0043231	intracellular membrane-bounded organelle**	0.001119	81	98,44	Down+Combined
98	GO:0005634	nucleus**	0.00419	46		Up+Combined
44	GO:0005739	mitochondrion*	0.045635	46		Down
20	GO:0005622	intracellular**	8.37E-04	72		Up+Down+Combined
6	GO:0032991	macromolecular complex**	2.22E-04	45	54	Down+Combined
54	GO:0043234	protein complex**	0.001672	6		Down+Combined

[a] 'Index' column is the index number used in the Figure 13. 'Term', 'Parent', 'Children' and 'Set_Desc' columns are defined as in Table 9.

Table 14. P-values for the number of total orthologs in each species for each set (Part 1) [a]

Set	Alarm_Pheromone	Old_vs_Young	Soldier_CG	Soldier_WG	Forager_CG	Forager_WG	Guard_CG	Guard_WG
S.cerevisiae	0.098665	1.74E-03	0.3585	0.5397	0.848	0.4154	0.9997	0.9997
T.adhaerens	0.750338	9.18E-02	0.36179	0.9268	0.9711	0.9966	0.9997	0.9994
N.vectensis	0.090393	2.16E-02	0.12939	0.8428	0.9945	0.9868	0.9948	1
S.mansoni	0.0632	0.688299	2.69E-02	0.8089	0.8677	0.7581	0.8939	0.9972
P.pacificus	0.326636	1.41E-02	0.64287	0.9176	0.9889	0.941	0.9924	0.9684
B.malayi	0.130638	0.019844	0.12118	0.9667	0.8609	0.9659	0.9998	0.9875
C.japonica	0.180326	6.36E-02	0.78364	0.7889	0.4888	0.8997	0.9995	0.9774
C.elegans	0.106166	0.341531	0.40439	0.8306	0.7633	0.9714	0.996	0.9821
C.brenneri	0.406177	0.108963	0.94366	0.9906	0.9353	0.98	0.9993	0.985
C.remanei	0.100921	0.239471	0.18648	0.7478	0.8292	0.9167	0.9916	0.9986
C.briggsae	0.0773	0.169211	0.54677	0.8568	0.7376	0.9959	0.9989	0.9853
L.gigantea	1.21E-02	0.023464	0.13258	0.5449	0.9581	0.998	0.9934	0.9871
C.spI	0.006568	1.12E-01	0.2581	0.8703	0.9707	0.9998	0.997	1
H.robusta	0.058051	0.407775	0.35757	0.9123	0.9776	0.999	1	0.9992
I.scapularis	0.084224	0.334674	0.82854	0.5021	0.7886	0.9553	0.9847	0.9964
D.pulex	1.06E-02	0.002097	0.00254	0.8897	0.9161	0.9994	0.8275	0.9919
P.humanus	1.92E-01	0.094586	3.92E-01	0.9596	0.9308	0.9445	0.8781	0.7908
A.pisum	0.14391	0.214726	0.0182	0.4207	0.1823	0.9683	0.6436	0.9289
N.vitripennis	0.158965	0.000192	0.04296	0.6392	0.4302	0.8204	0.0276	0.7315
A.mellifera	1	1	1	1	1	1	1	1
T.castaneum	4.26E-02	0.000618	0.3152	0.5712	0.5135	0.8662	######	0.8609
B.mori	1.84E-02	0.005909	1.25E-02	0.7735	0.4778	0.9476	0.1889	0.9605
A.gambiae	1.04E-01	0.000726	0.00015	0.1455	0.2712	0.8093	######	0.9949
A.aegypti	0.32798	0.012738	0.00062	0.4637	0.6281	0.3623	0.5266	0.9991
C.pipiens	5.92E-02	0.048128	4.10E-05	0.5691	0.2293	0.8981	######	0.9375
D.grimshawi	0.011206	0.000644	2.00E-06	0.4628	0.537	0.9346	0.2476	0.9402
D.virilis	5.26E-02	0.009588	2.80E-05	0.6453	0.5876	0.9332	0.4412	0.8752
D.mojavensis	4.93E-02	0.00154	1.00E-06	0.5339	0.3184	0.8524	0.3216	0.9337
D.willistoni	0.334092	0.002795	0.00064	0.6209	0.3098	0.8508	0.5888	0.9021
D.pseudo-obscura	3.14E-02	0.000407	0.00123	0.6427	0.879	0.7041	0.3008	0.8361
D.ananassae	7.04E-02	0.001816	1.90E-05	0.5902	0.5448	0.8451	0.3375	0.8928
D.melano-gaster	7.09E-02	0.007964	0.00011	0.6278	0.4618	0.8815	0.2732	0.9273

[a] Raw data for Figure 11, p-values are calculated by random sampling.

Table 15. P-values for the number of total orthologs in each species for each set (Part 2) [a]

Set	Alarm_Pheromone	Old_vs_Young	Soldier_CG	Soldier_WG	Forager_CG	Forager_WG	Guard_CG	Guard_WG
S.purpuratus	2.75E-02	1.10E-01	3.09E-02	0.8392	0.9288	0.9733	0.9964	0.991
C.savignyi	0.623682	2.66E-02	0.86674	0.9206	0.9547	0.9644	0.9998	0.9905
C.intestinalis	0.417536	3.30E-02	0.98303	0.883	0.9804	0.9917	0.9971	0.9831
B.floridae	2.13E-02	4.98E-02	4.50E-02	0.7568	0.9668	0.9964	0.9923	0.9993
T.nigroviridis	0.287014	2.80E-01	1.33E-01	0.9009	0.9204	0.9988	0.9992	0.9924
T.rubripes	1.95E-02	2.27E-01	8.19E-02	0.9681	0.9447	0.9967	0.9995	0.998
G.aculeatus	0.007768	0.087616	1.66E-01	0.6059	0.6756	0.9744	0.9983	0.9951
O.latipes	0.195762	0.031929	1.06E-01	0.8389	0.7927	0.983	0.993	0.9961
D.rerio	1.16E-01	0.046795	1.53E-01	0.7564	0.9802	0.9318	0.9889	0.9942
X.tropicalis	2.18E-02	0.132188	0.67786	0.9439	0.9645	0.9735	0.9997	0.997
G.gallus	0.083907	0.678986	0.27258	0.7935	0.7398	0.9961	0.9989	0.9916
O.anatinus	0.015797	0.340381	0.7995	0.7683	0.8227	0.9437	0.9993	0.9874
M.domestica	7.52E-02	0.103052	0.25541	0.8738	0.891	0.997	0.9998	0.9992
B.taurus	1.22E-02	3.85E-01	0.00587	0.6966	0.9348	0.997	0.9995	0.9924
E.caballus	0.011487	1.70E-01	1.56E-01	0.9226	0.8218	0.9965	0.9999	1
C.familiaris	3.83E-02	0.17433	1.68E-01	0.8864	0.9588	0.9972	0.9996	0.9962
C.porcellus	3.36E-02	0.092751	7.47E-02	0.8518	0.909	0.9994	0.9997	0.9999
R.norvegicus	0.003046	0.194172	0.01559	0.6537	0.8467	0.9876	0.9937	0.9591
M.musculus	0.004545	0.203356	0.01336	0.8137	0.9137	0.9982	0.998	0.9993
M.mulatta	0.007764	0.172438	1.29E-01	0.884	0.9082	0.997	0.9993	0.9951
P.pygmaeus	9.63E-03	0.274949	5.64E-02	0.8168	0.9714	0.989	0.9999	0.978
P.troglodytes	1.58E-02	0.104355	0.01752	0.8426	0.9393	0.9976	1	0.9952
H.sapiens	0.013503	0.245454	7.42E-02	0.9313	0.9761	0.9935	1	0.974

[a] Raw data for Figure 11, p-values are calculated by random sampling.

Table 16. P-values for pairwise comparison of ortholog numbers of different gene sets using parametric/non-parametric test [a]

Kolmogorov-Sirnov Test	AlarmPheromone	OldvsYoung	SoldierCG	SoldierWG	ForagerCG	ForagerWG	GuardCG	GuardWG
AlarmPheromone	1	0.367123	0.714213	**0.0251346**	**0.0282667**	**0.00248144**	**0.00109826**	**5.34150e-4**
OldvsYoung		1	0.134597	0.152380	0.0641551	**0.0136671**	**0.00348467**	**2.08800e-4**
SoldierCG			1	**0.0258303**	0.0633294	**0.00542168**	**0.00178655**	**6.51500e-4**
SoldierWG				1	0.916879	0.0519380	0.229496	**0.0289427**
ForagerCG					1	0.122075	0.946011	0.235732
ForagerWG						1	0.192456	0.325085
GuardCG							1	0.611793
GuardWG								1
T Test	AlarmPheromone	OldvsYoung	SoldierCG	SoldierWG	ForagerCG	ForagerWG	GuardCG	GuardWG
AlarmPheromone	1	0.349017	0.490560	**0.0188484**	**0.0122057**	**1.62244e-3**	**0.00026134**	**5.19858e-05**
OldvsYoung		1	0.783316	**0.0466855**	**0.03475844**	**0.0041887**	**0.00045919**	**0.00012456**
SoldierCG			1	**0.0374450**	**0.02460937**	**2.23987e-3**	**0.00036474**	**7.00051e-05**
SoldierWG				1	0.68143143	**0.0356969**	0.174657	**0.03606973**
ForagerCG					1	0.0517752	0.420649	0.107004
ForagerWG						1	0.100734	0.263167
GuardCG							1	0.321718
GuardWG								1

[a] The 8 gene sets here are the same sets that are described in Table 6. For each gene set, the set of the number of orthologs of each gene is used for non-parametric Kolmogorov-Smirnov test (KS-test), the 'Average Number of Orthologs per gene' and 'StandardDeviation' in Table 7 are used for parametric Two-sample T Test. Both The Kolmogorov-Smirnov test (KS-test) and the Two-sample T Test are two-tail test here, p-value smaller than 0.05 are bolded.

Table 17. D2z hits of 'Vertebrate-conserved' Alarm_Pheromone Genes that are related to neural disorders [a]

Entrez ID	No of probes picking it	Name	Neuro-related-Diseases	References PubMed IDs
3241	2	HPCAL1	alzheimer's disease	PMID: 21059989
7227	5	TRPS1	Langer-giedion syndrome	PMID: 8530105
146691	5	TOM1L2	dementia	PMID: 20167577
885	2	CCK	panic disorder	PMID: 8185172
10385	2	BTN2A2	schizophrenia	PMID: 19785721
23446	2	SLC44A1	neuroblastoma	PMID:19519661
57508	3	INTS2	seizures	PMID: 10390770
166614	2	DCLK2	lissencephaly;quadriplegia;seizures	PMID: 17997185
6925	3	TCF4	Pitt-hopkins syndrome	PMID: 17436255
143279	11	HECTD2	prion disease ;alzheimer's disease	PMID: 19214206; PMID: 19754925
23295	2	MGRN1	neurodegeneration	PMID: 17720281
65250	2	C5orf42	joubert syndrome;monomelic amyotrophy	PMID: 22425360;PMID: 22264561
25939	3	SAMHD1	Aicardi-goutieres syndrome	PMID: 19525956
51473	3	DCDC2	dyslexia ;Attention Deficit Hyperactivity Disorder	PMID: 16385449
56853	2	CELF4	rontotemporal dementia	PMID: 15009664

[a] 'Vertebrate-conserved' here means the Alarm_Pheromone genes have orthologs in all vertebrate species of InParanoid (altogether 19, ranging from *T.nigroviridis* to *H.sapiens*).Those neural-disorder-related hits that are specifically mentioned in the text are not listed here.

Figures

Figure 1. Distribution of Genes in human 7q36.1a chromosome cluster across species. Maps of the human 7q36.1a gene cluster in human chromosome 9 (hg19 sequence build, chr 7), dog (canfFam3) chr 15, mouse (mm9) chr6, opossum (monDom5) chr8, and the fragmented genomes of chicken (galGal3), zebra finch (taeGut1) and lizard (AnoCar 2.0) are shown. Colored dotted lines connect orthologs in the different species. chrUn is assigned to genes in fragmented assemblies that have not been assigned to specific chromosomes in some species. ZNF786, ZNF425, ZNF398, ZNf212, and ZNF746 are HUB- and KRAB-containing ZNF genes that are closely related to ZNF282, ZNF777 and ZNF783. ZNF467, a deeply conserved "ZNF only" gene that is also found clustered with the KRAB-ZNF genes in mammals, is also shown.

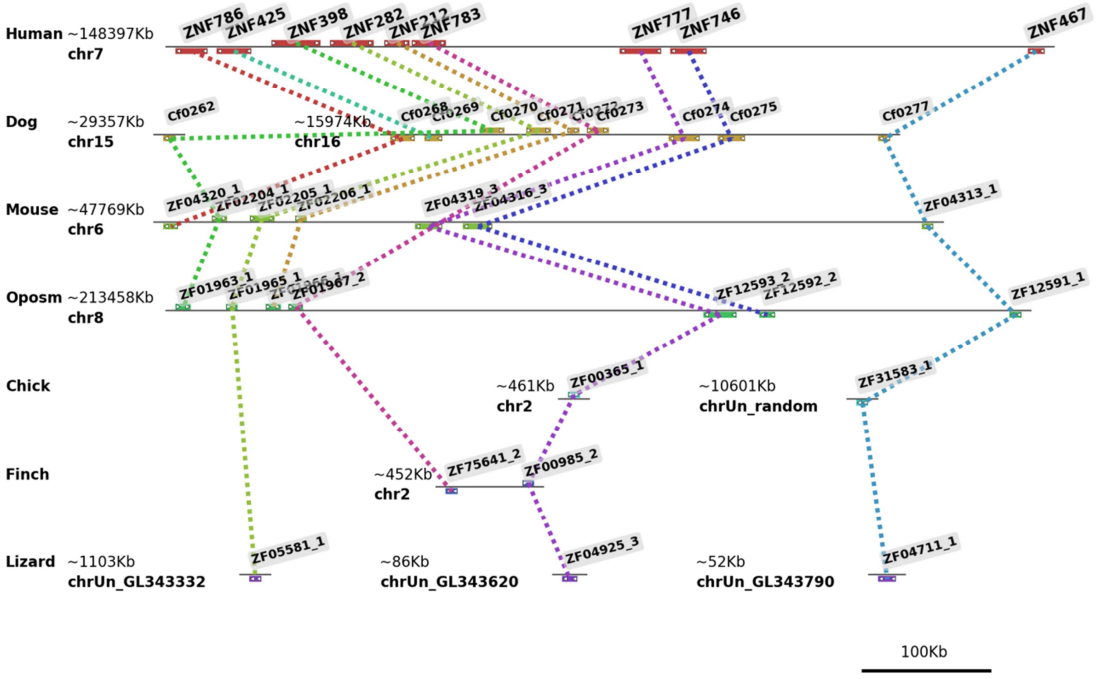


Figure 2. Evolutionary tree showing relationships between KRAB domain sequences from human, opossum, chicken and frog. Left is a circular tree showing clustering of sequences including all ZNF-associated KRAB domains from human (red), opossum (purple), chicken (blue) and frog (green) KRAB-ZNF gene models (see Methods). The boxed region is expanded and shown as a rectangular tree on the right. The KRAB domain of the S. purpuratus PRDM9 protein was included to root the tree. The human ZNF282-related cluster dominates this branch of the tree which also includes isoforms of ZNF862, a KRAB-containing TTF-finger zinc finger gene that also maps to the cluster region in human chromosome 7q36.1.

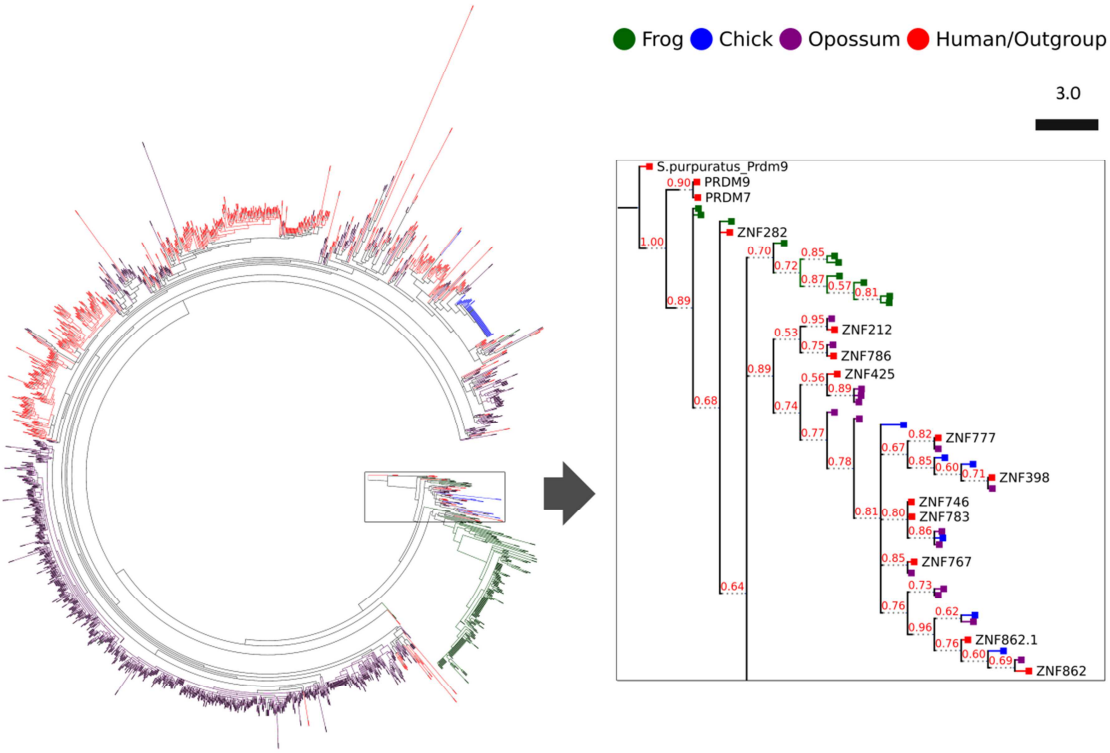


Figure 3. Consensus sequences of KRAB domains from human, opossum, chicken, frog and krab genes in 7q36 cluster. Numbers under the X-axis in each panel represent amino acid positions, in N- to Cterminal orientation, in the consensus derived from all ZNF-linked KRAB domain sequences in each species; the y axis represents information content (bits) at each position. The height of each letter represents the frequency with which amino acids represented by the letters are found at each position. Asterisks above certain letters at each position indicate agreement with the sequence that has been determined to be necessary for KAP1 binding in human KRAB sequences, at positions 6,7 (DV) and 36-38 (MLE). Consensus sequence of 7q36.1 is generated by Krab_A sequences of ZNF786, ZNF425, ZNF398, ZNF212, ZNF783, ZNF777, ZNF746 of human, mouse, opossum, lizard and finch.

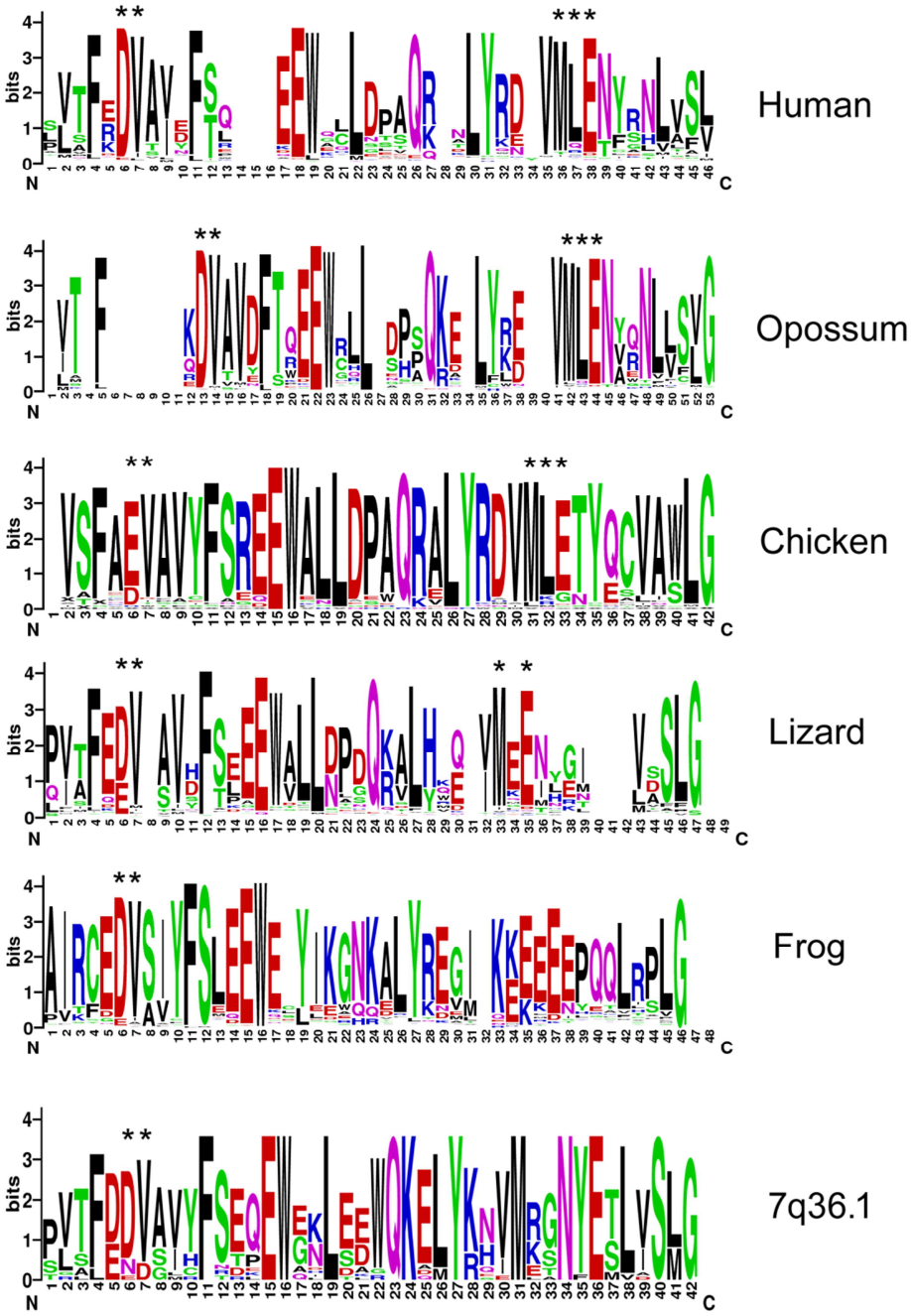


Figure 4. RNA *in situ* hybridization in sectioned mouse (A) and chicken (B) embryos at embryonic day 12.5 (E12.5) or 7 (d7), respectively. Despite differences in the pace of development of brain compared to other tissues in these specials, expression in forebrain (fb), hindbrain (hc) ganglia (ga), spinal cord (sc) and in the developing structures of the face is very similar in the two species.

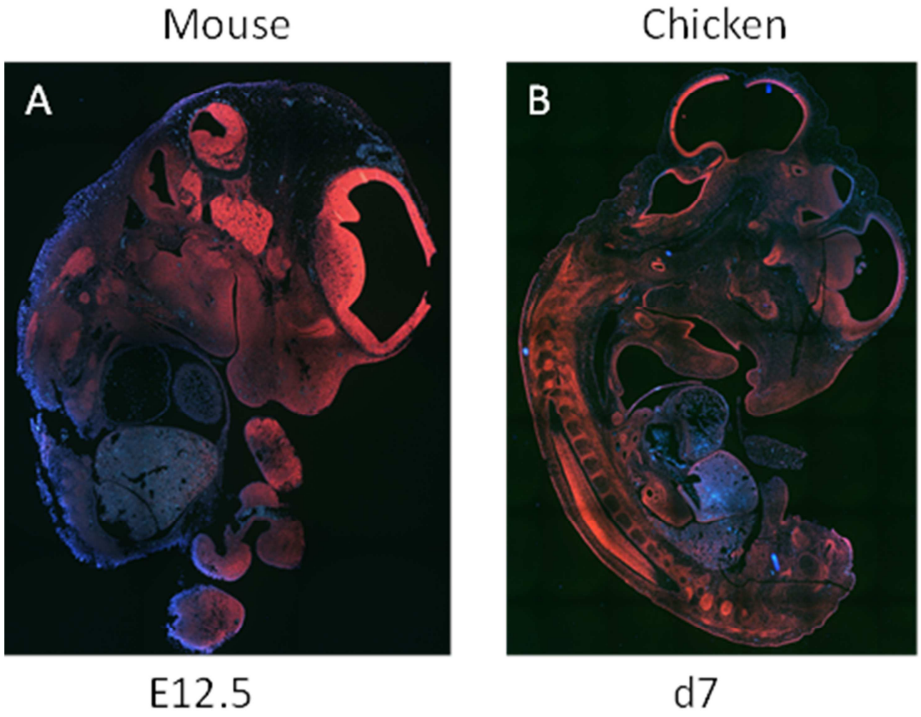


Figure 5. RNA-seq expression patterns for deeply conserved (A) or primate-specific (B) polydactyl ZNF genes in adult human tissues. For both groups,expression is especially high in reproductive and immune tissues. Arrow to the left of panel (A) is between the adjacent positions of ZNF282 and ZNF777, which are tightly clustered in their expression. Expanded versions of both panels with gene names associated are provided in supplementary materials of the paper[164]. Sk.Muscle: skeletal muscle; Wh.BloodCell: white blood cells.

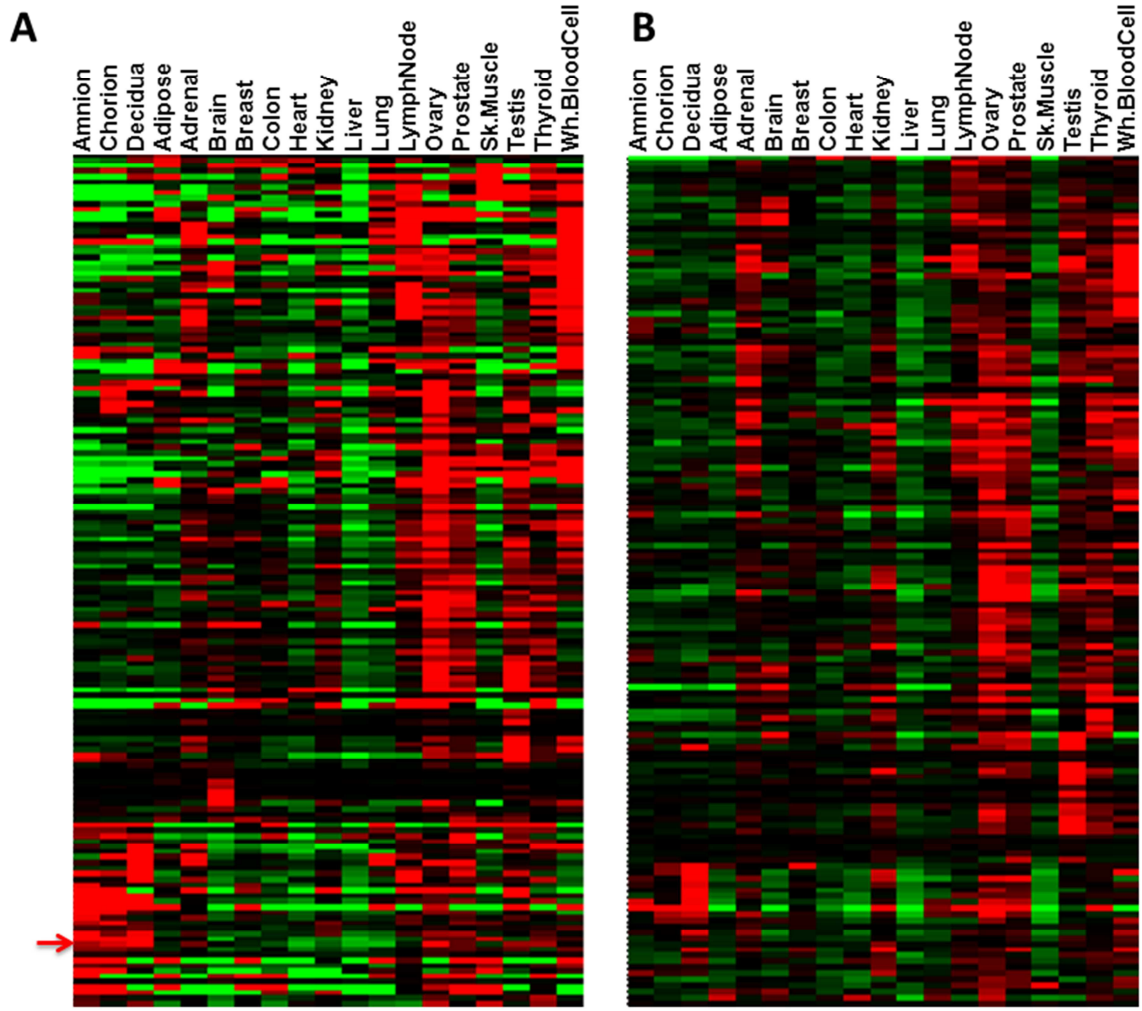


Figure 6. Expression of Zfp282 and Zfp777 genes in embryonic mouse tissues and ZNF282 and ZNF777 proteins in human term placenta. Mouse Zfp777(A) and Zfp282 (B) transcriptswere measured in RNA extracted from dissected embryos and extraembryonic tissues isolated from embryonic day 12.5 (E12.5) to E18.5. Relative expression, measured against the average levels of two ubiquitous genes, Sdha and Ywhaz, are plotted across 2-day gestational intervals in decidua (blue bars), fetal placenta (red), yolk sac (green), fetal heads (purple), fetal bodies (turquoise), and fetal liver (gold). Antibodies specific to the human ZNF777 and ZNF282 proteins (stained in red) were also used to track cell-type specific protein expression in sectioned human term placenta. The sections were counterstained with Hoechst dye (blue) to highlight locations of nuclei. Panels (C) (ZNF777) and (D) (ZNF282) show lower resolution views of placental regions near the maternal:fetal interface including maternal decidual cells (dc), fetal anchoring villi (av), and floating villi (fv) surrounding maternal blood spaces (mbs). An arrowhead in panel (D) highlights the location of a maternal lymphocyte (lc) that is brightly stained by the ZNF282 antibody. Panels (E–G) show higher magnification images from the ZNF777 IHC highlighting decidual cells (dc, panel E) and fetal syncytiotrophoblasts (st) lining anchoring (F) and floating villi (G). White bar in each image represents 25

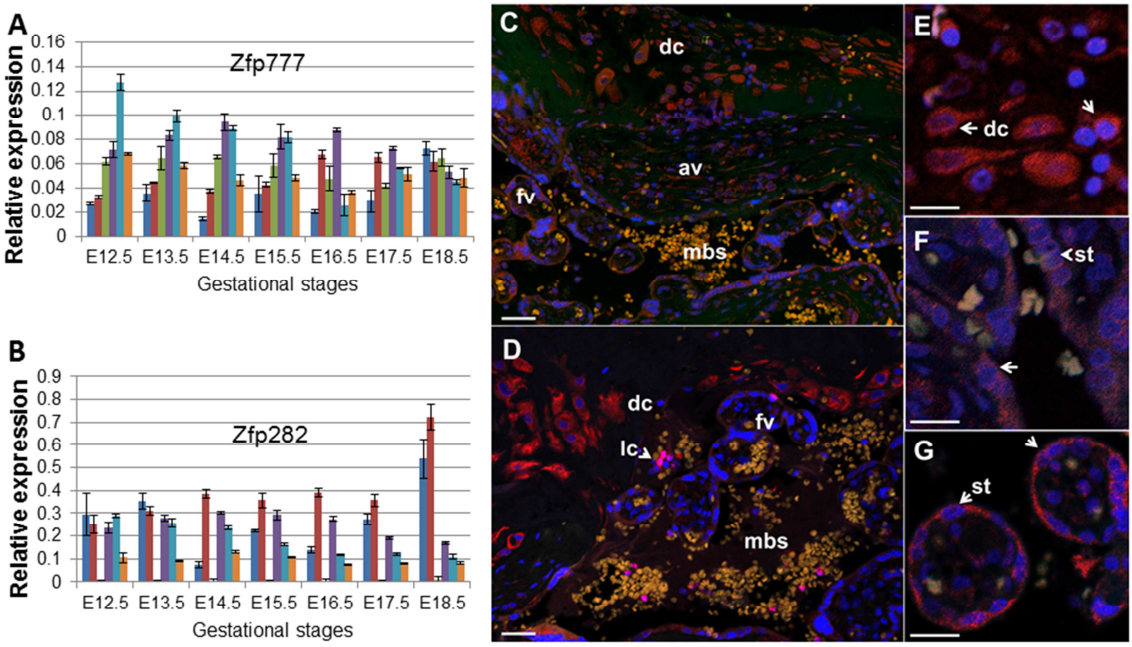


Figure 7. Distribution of Average Ortholog Number for Forager_CG set. This distribution is generated by random sampling of 1 million times. Each time, a random set of the same size of Forager_CG set is retrieved from InParanoid's whole honey bee gene population (that are on the oligo array as defined in Methods). And the Average Ortholog Number of the random set is calculated: [the total number of orthologs of all genes within the set]/[number of genes within the set].

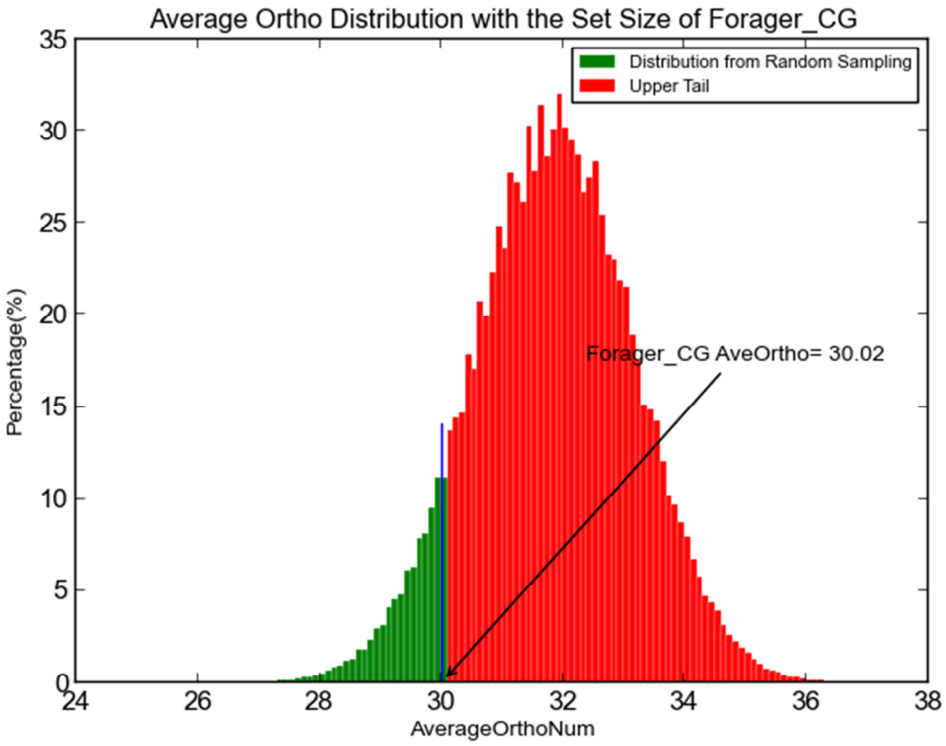


Figure 8. Distribution of Average Ortholog Number for Alarm_pheromone set. This distribution is generated by random sampling of 1 million times. Each time, a random set of the same size of Alarm_pheromone set is retrieved from InParanoid's whole honey bee gene population (that are on the oligo array as defined in Methods). And the Average Ortholog Number of the random set is calculated: [the total number of orthologs of all genes within the set]/[number of genes within the set]

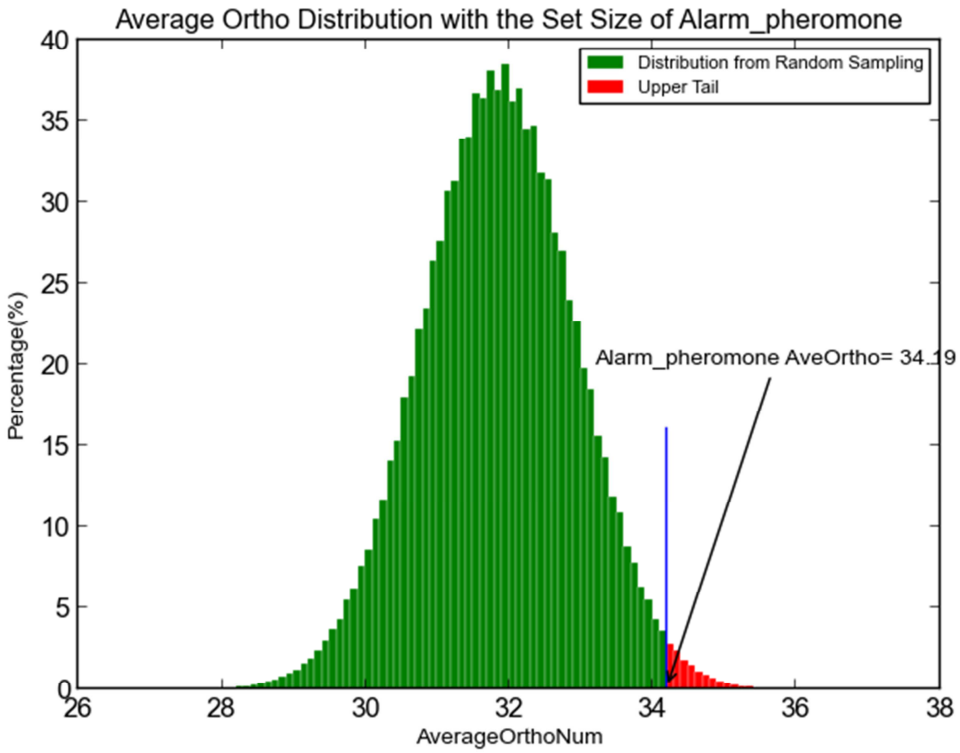


Figure 9. Normalized distribution of the number of all "Array-spotted" Honey Bee genes' orthologs in 54 species. "Array-spotted" means that these genes are present in the InParanoid database and spotted on the Honey Bee Oligonucleotide Microarray. There are 7462 such honey bee genes. X-axis is the number of orthologs in 54 species (53 metazoan species+ yeast). Y-axis is the percentage of these 7462 honey bee genes that have the corresponding number of orthologs.

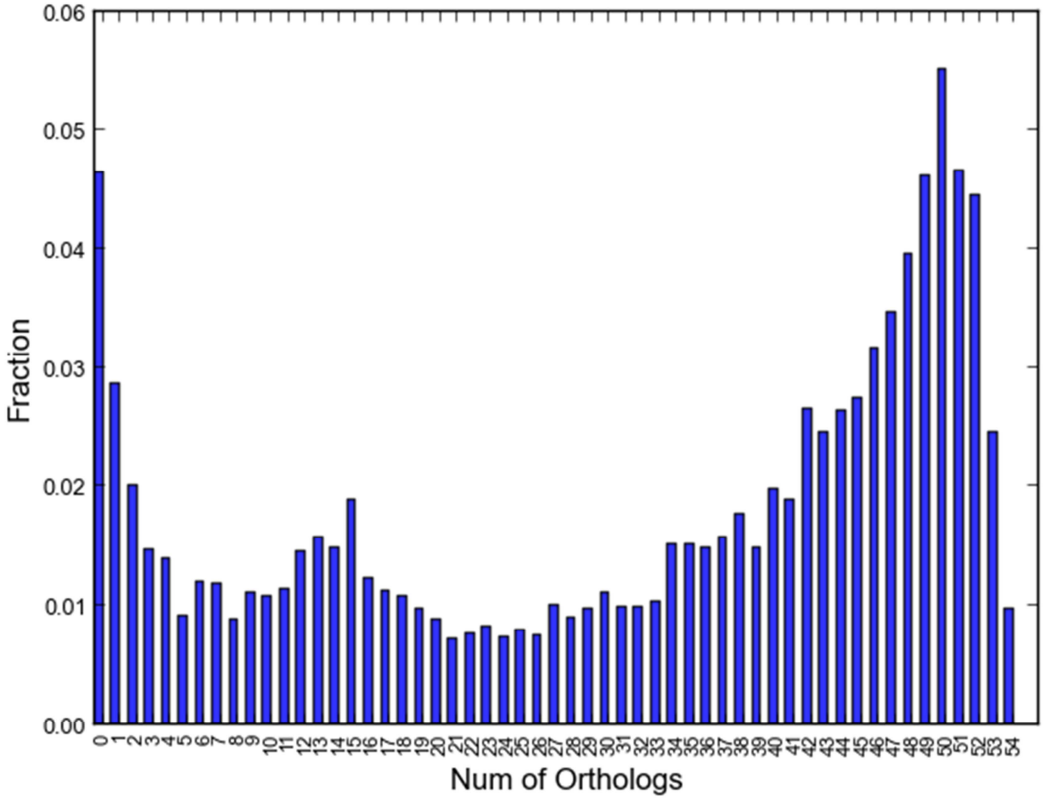


Figure 10. Normalized distribution of the number of all "Array-unspotted" Honey Bee genes' orthologs in 54 species. "Array-unspotted" means that these genes are present in the InParanoid database but not spotted on the Honey Bee Oligonucleotide Microarray. There are 1631 such honey bee genes. X-axis is the number of orthologs in 54 species (53 metazoan species+ yeast). Y-axis is the percentage of these 1631 honey bee genes that have the corresponding number of orthologs. (Note vertical scale difference between Figure 9 and 10)

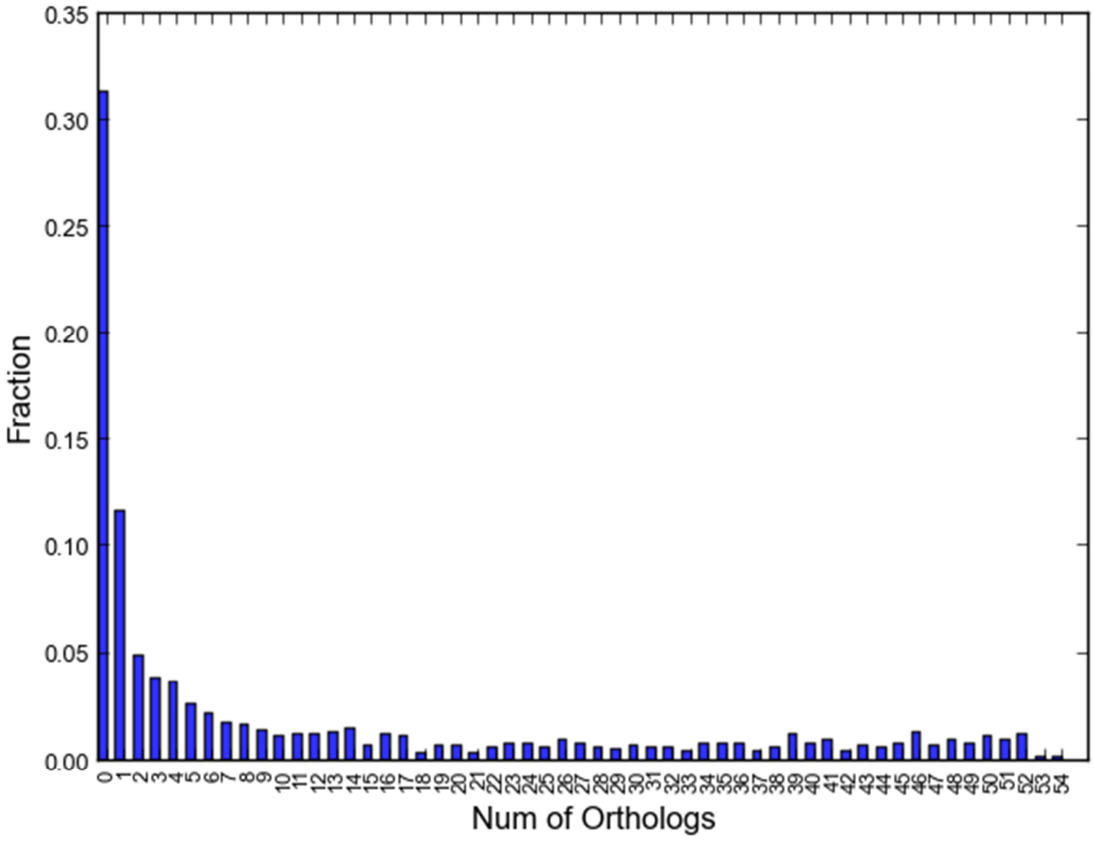


Figure 11. Heat map of the p-values of number of orthologs for each species and each set. Each row of the map represents one of the honey bee experimental sets. Each column represents one species. One column that is completely white represents the honey bee. The species are ordered along the x-axis by evolutionary distance from the honey bee based on NCBI taxonomy common tree (http://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi) of the 55 species. The order is further refined based on the tree from Flybase (http://flybase.org/blast/species_tree.png), WormBook[165] and UCSC genome browser[166]. Those species to the left of the honey bee (the white vertical column) belong to lineages that diverged from the insects earlier than the insects diverged from the lineage leading to the mammals. Yeast is the leftmost species. The 12 species immediately to the right of the honey bee are insects. The ten species at the far right are placental mammals, with H. sapiens being the farthest to the right. Between the insects and the placental mammals are marine invertebrates, marine chordates, fish, amphibians, birds, and one marsupial, the short-tailed gray opossum. The color code (vertical bar on the right hand side of the figure) represents the p-value for statistical significance of enhancement of orthology in each species relative to the degree of orthology of all the genes on the microarray. Grayscale from black to white indicates a range of p-values from .05 to 1, with the black being very close to .05 and white being very close to 1.0. Color scale is shown for p-value 0 to .05, with a red-yellow mix, so pure deep red is very near zero and yellow is very near .05. P-values above 0.5 indicate relative depletion in conserved proteins while p-values below 0.5 indicate relative enhancement. The numerical data for this plot is in Table 14,15.

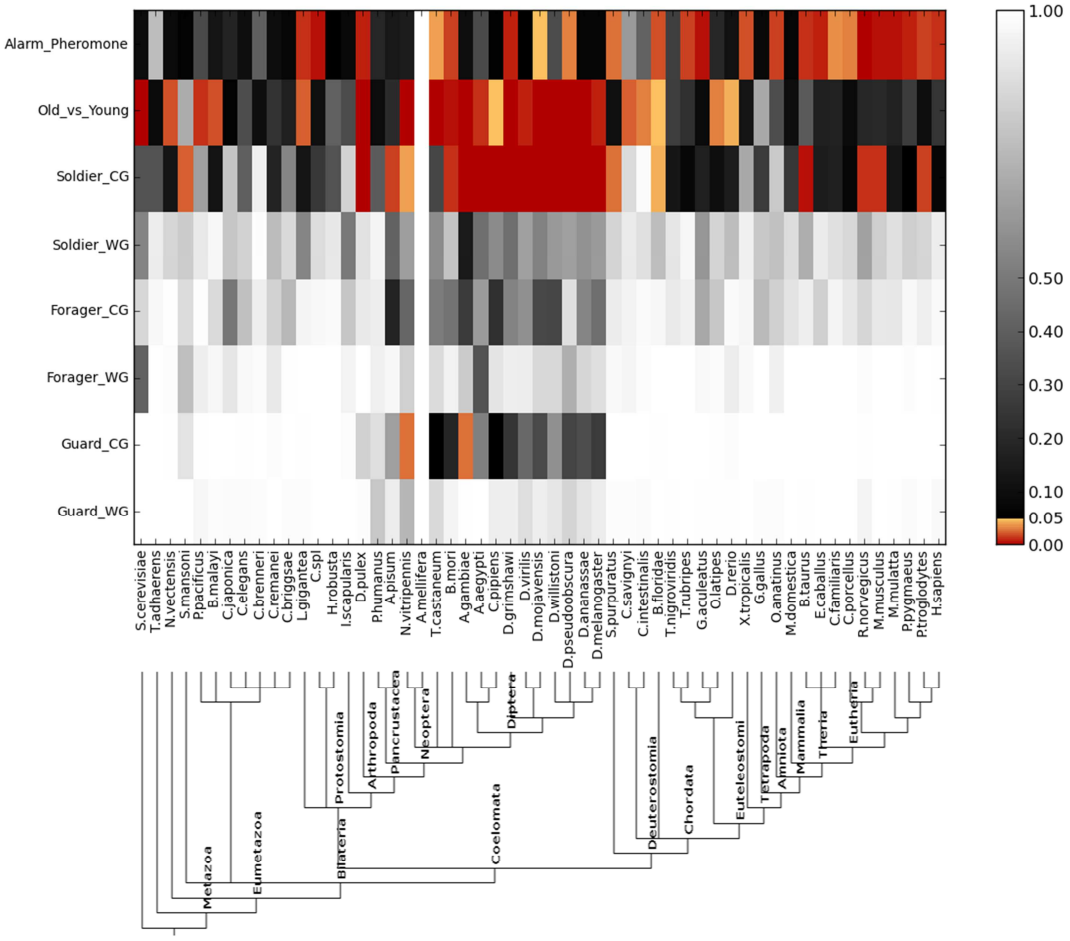


Figure 12. GO Trees (p-value cutoff = 0.05) for the 'Vertebrate-conserved' Alarm_Pheromone Genes' Human Orthologs(Part 1). This is the Biological Process Tree. 'Vertebrate-conserved' means that genes have orthologs in all vertebrate species included in InParanoid database. Nodes with red bounds are GO terms with p-value <= 0.01. Nodes with various green bounds are nodes with p-value between 0.01 and 0.05. Nodes with white bounds (or no bounds as it is the same as the background color) are not themselves enriched but are parents of enriched terms in the GO hierarchy. Separate GO analysis are done for all the up-regulated 'Vertebrate-conserved' Alarm_Pheromone Genes' Human orthologs ('Up'),all the down-regulated ones('Down'), and all no matter up or down-regulated('Combined'). The results are indicated by the colors of the node faces as follows: GO category enriched in the up-regulated subset only is red; GO category enriched in the down-regulated subset only is green; GO category enriched in the complete set is deep blue; GO category enriched in the up-regulated subset and the complete set is purple; GO category enriched in the down-regulated subset and the complete set is light blue; GO category enriched in both subsets and in the total set is gray. For example, GO term indexed '13' is showed up as significant in the analysis for down-regulated and for all genes (Hence it is in light blue, 'Down+Combined'). *There are no terms significant in 'Up+Down' category, so this category is not shown in the color coding.* The exact names for these indexed GO terms are in Table 9, 10 and 11(The tables and this tree are designed to be complementary to each other. The trees show the overall architecture of the relationships among GO categories while the tables provide more detail) or Appendix A Table A13.

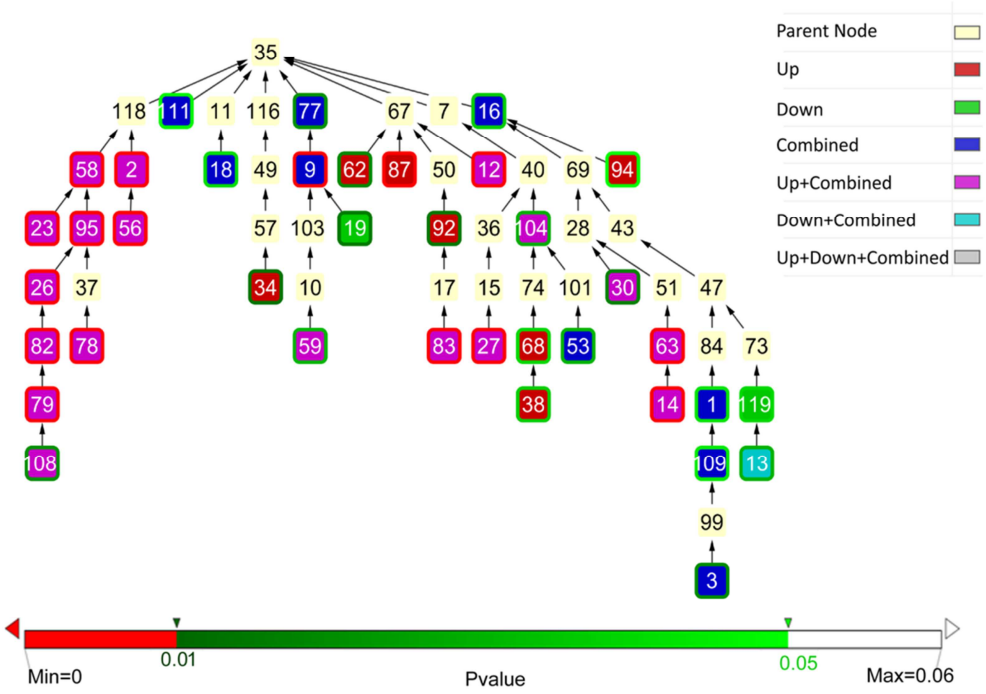


Figure 13. GO Trees (p-value cutoff = 0.05) for the 'Vertebrate-conserved' Alarm_Pheromone Genes' Human Orthologs(Part 2). These are the Molecular Function (left) and Cellular Component (right) trees. 'Vertebrate-conserved' defined as in Figure 12 legend. Nodes with red bounds are GO terms with p-value <= 0.01. Nodes with red bounds are GO terms with p-value <= 0.01. Nodes with various green bounds are nodes with p-value between 0.01 and 0.05. Nodes with white bounds (or no bounds as it is the same as the background color) are not themselves enriched but are parents of enriched terms in the GO hierarchy. Separate GO analysis are done for all the up-regulated 'Vertebrate-conserved' Alarm_Pheromone Genes' Human orthologs ('Up'),all the down-regulated ones('Down'), and all no matter up or down-regulated('Combined'). The results are indicated by the colors of the node faces as follows: GO category enriched in the up-regulated subset only is red; GO category enriched in the down-regulated subset only is green; GO category enriched in the complete set is deep blue; GO category enriched in the up-regulated subset and the complete set is purple; GO category enriched in the down-regulated subset and the complete set is light blue; GO category enriched in both subsets and in the total set is gray.. The exact names for these indexed GO terms are in Table 12,13 or Appendix A Table A13.

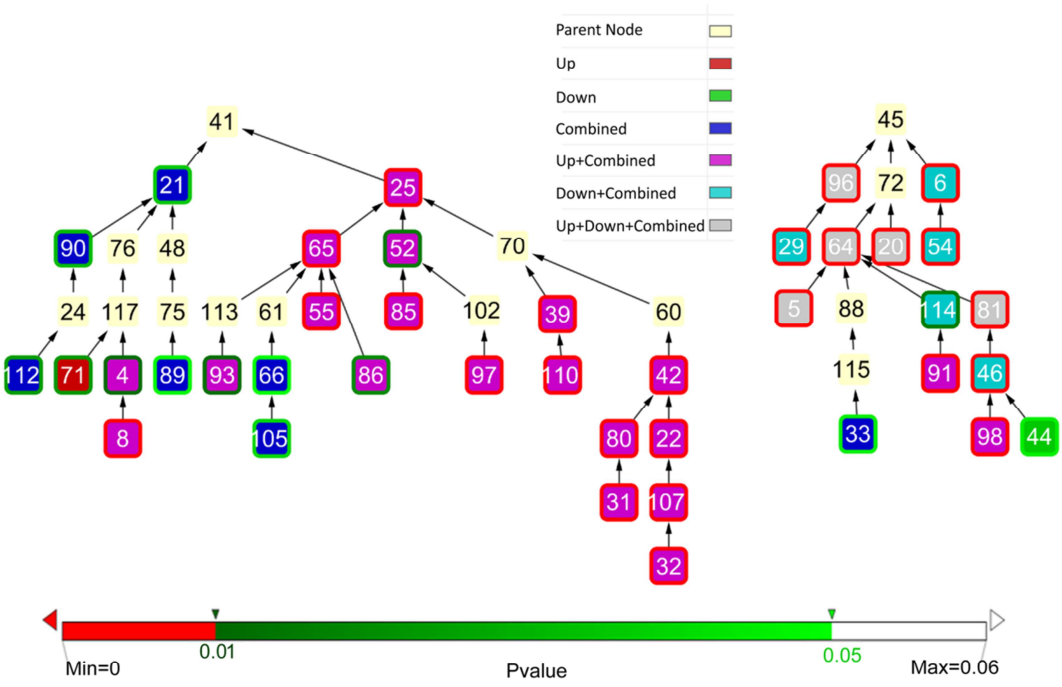


Figure 14. GO Biological Process Tree (p-value cutoff = 0.01) for the 'Mouse-human-conserved' Alarm_Pheromone Genes' Human Orthologs. 'Mouse-human-conserved' means that genes have orthologs in both mouse and human. Nodes with red bounds are GO terms with p-value <= 0.01. Nodes with white bounds are GO terms that have p-value > 0.01 (i.e. are just nodes that are parents of the significant GO terms in the GO hierarchy). Separate GO analysis are done for all the up-regulated 'Mouse-human-conserved' Alarm_Pheromone Genes' Human orthologs ('Up'),all the down-regulated ones('Down'), and both up or down-regulated('Combined'). For example, GO term indexed '234' is showed up as significant in the analysis for down-regulated and for all genes (Hence it is in light blue,'Down+Combined'). The exact names for these indexed GO terms are in Appendix A Table A13.

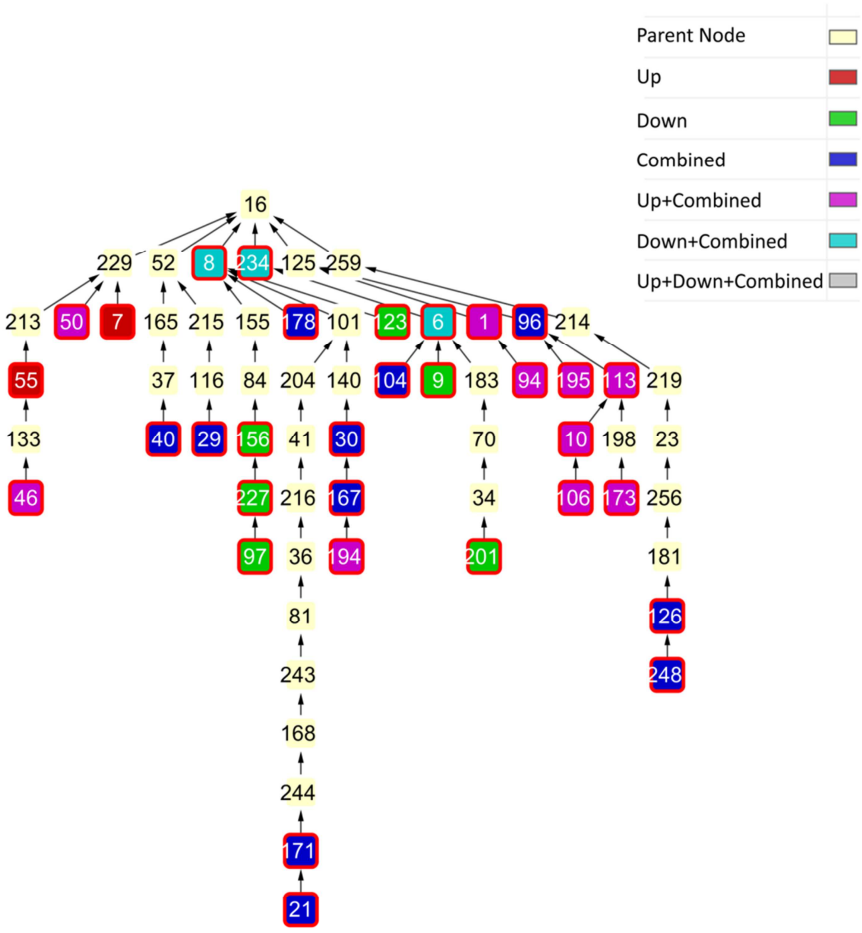


Figure 15. GO Molecular Function Tree (p-value cutoff = 0.01) for the 'Mouse-human-conserved' Alarm_Pheromone Genes' Human Orthologs. 'Mouse-human-conserved' means that genes have orthologs in both mouse and human. Nodes with red bounds are GO terms with p-value <= 0.01. Nodes with white bounds are GO terms that have p-value > 0.01 (i.e. are just nodes that are parents of the significant GO terms in the GO hierarchy). Color coding of the face and the exact names of nodes are the same as explained in Figure 14.

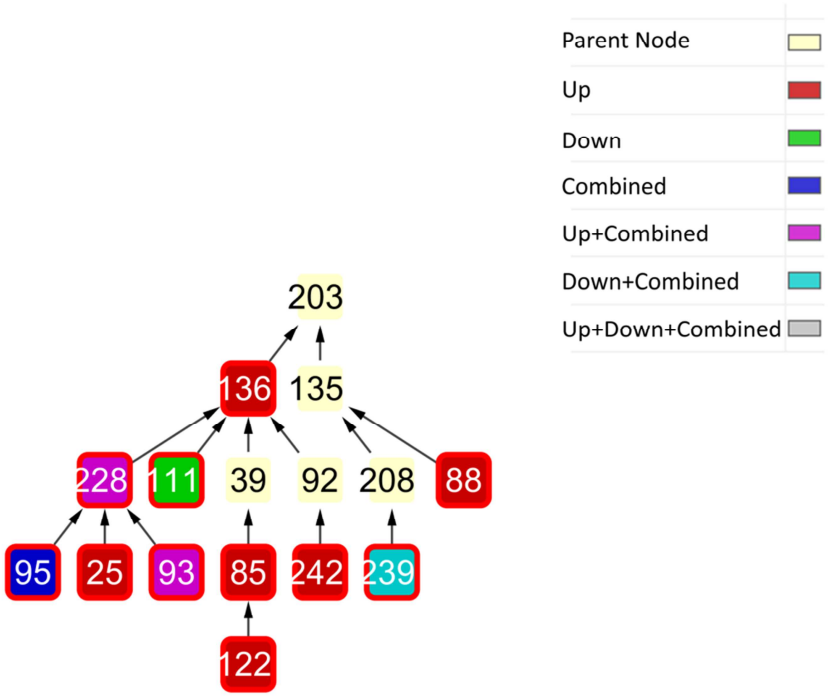


Figure 16. GO Cellular Component Tree (p-value cutoff = 0.01) for the 'Mouse-human-conserved' Alarm_Pheromone Genes' Human Orthologs. 'Mouse-human-conserved' means that genes have orthologs in both mouse and human. Nodes with red bounds are GO terms with p-value <= 0.01. Nodes with white bounds are GO terms that have p-value > 0.01 (i.e. are just nodes that are parents of the significant GO terms in the GO hierarchy). Color coding of the face and the exact names of nodes are the same as explained in Figure 14.

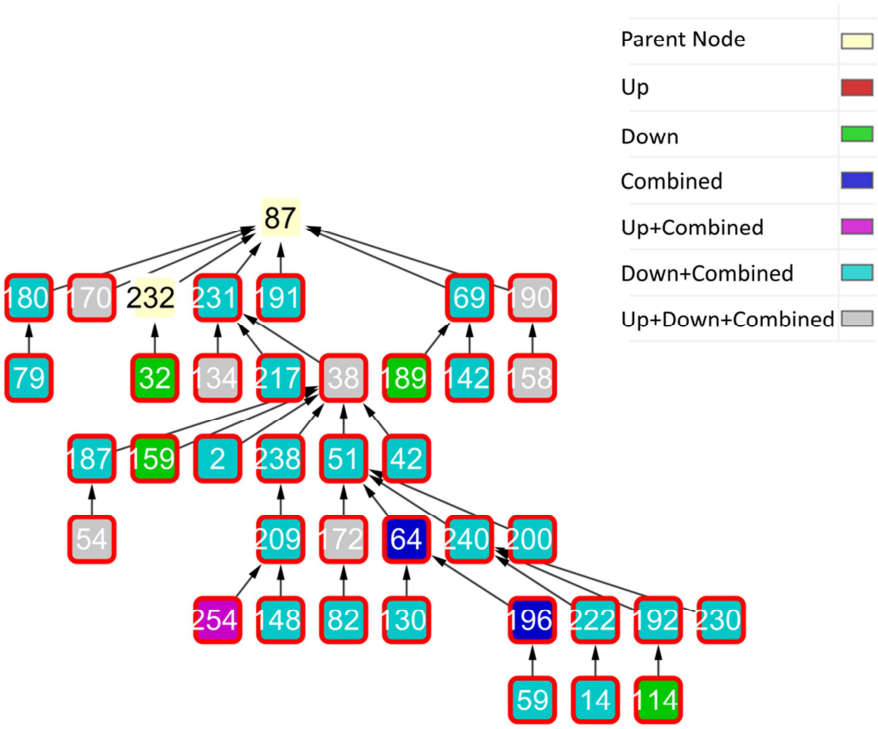


Figure 17. GO Biological Process Tree (p-value cutoff = 0.05) for the 'Mouse-human-conserved' Alarm_Pheromone Genes' Human Orthologs. 'Mouse-human-conserved' means that genes have orthologs in both mouse and human. Nodes with red bounds are GO terms with p-value <= 0.01. Nodes with green bounds are GO terms with p-value >0.01 and <=0.05. Nodes with white bounds are GO terms that have p-value > 0.05 (i.e. or are just nodes that are parents of the significant GO terms in the GO hierarchy). Color coding of the face and the exact names of nodes are the same as explained in Figure 14.

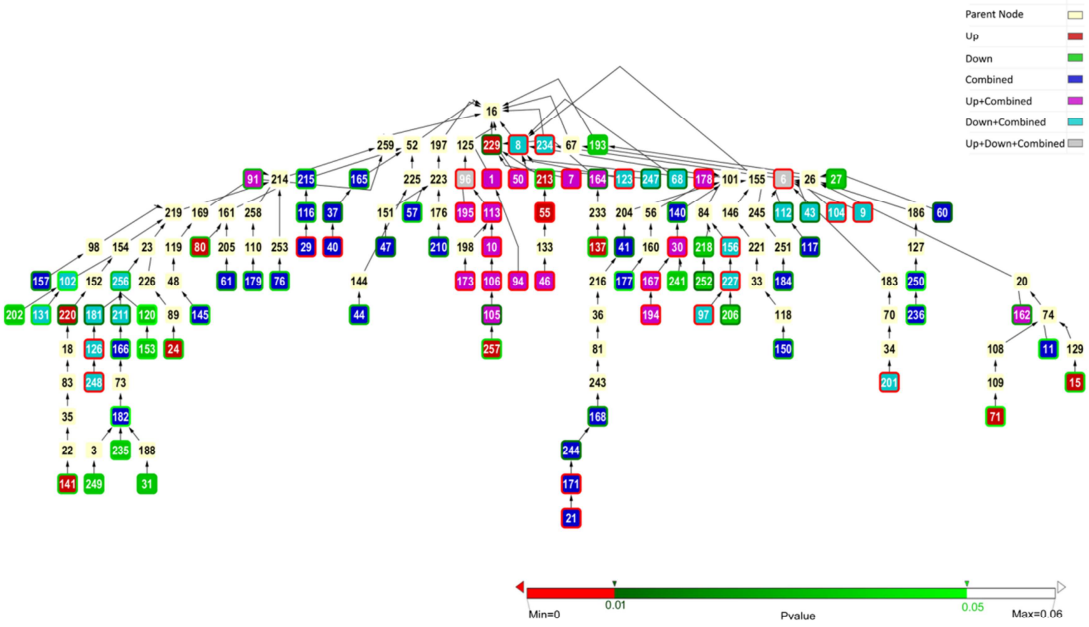


Figure 18. GO Molecular Function Tree (p-value cutoff = 0.05) for the 'Mouse-human-conserved' Alarm_Pheromone Genes' Human Orthologs. 'Mouse-human-conserved' means that genes have orthologs in both mouse and human. Nodes with red bounds are GO terms with p-value <= 0.01. Nodes with green bounds are GO terms with p-value >0.01 and <=0.05. Nodes with white bounds are GO terms that have p-value > 0.05 (i.e. or are just nodes that are parents of the significant GO terms in the GO hierarchy). Color coding of the face and the exact names of nodes are the same as explained in Figure 14.

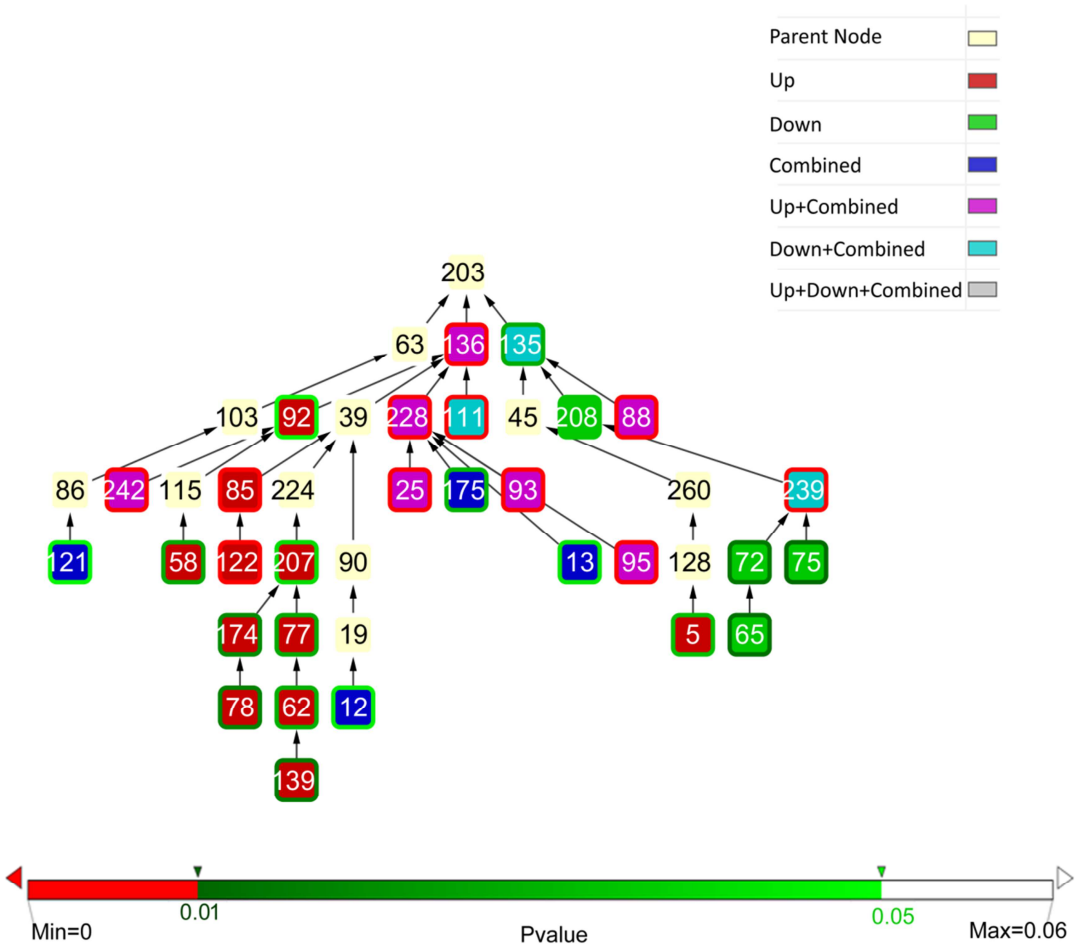


Figure 19. GO Cellular Component Tree (p-value cutoff = 0.05) for the 'Mouse-human-conserved' Alarm_Pheromone Genes' Human Orthologs. 'Mouse-human-conserved' means that genes have orthologs in both mouse and human. Nodes with red bounds are GO terms with p-value <= 0.01. Nodes with green bounds are GO terms with p-value >0.01 and <=0.05. Nodes with white bounds are GO terms that have p-value > 0. 05 (i.e. or are just nodes that are parents of the significant GO terms in the GO hierarchy). Color coding of the face and the exact names of nodes are the same as explained in Figure 14.

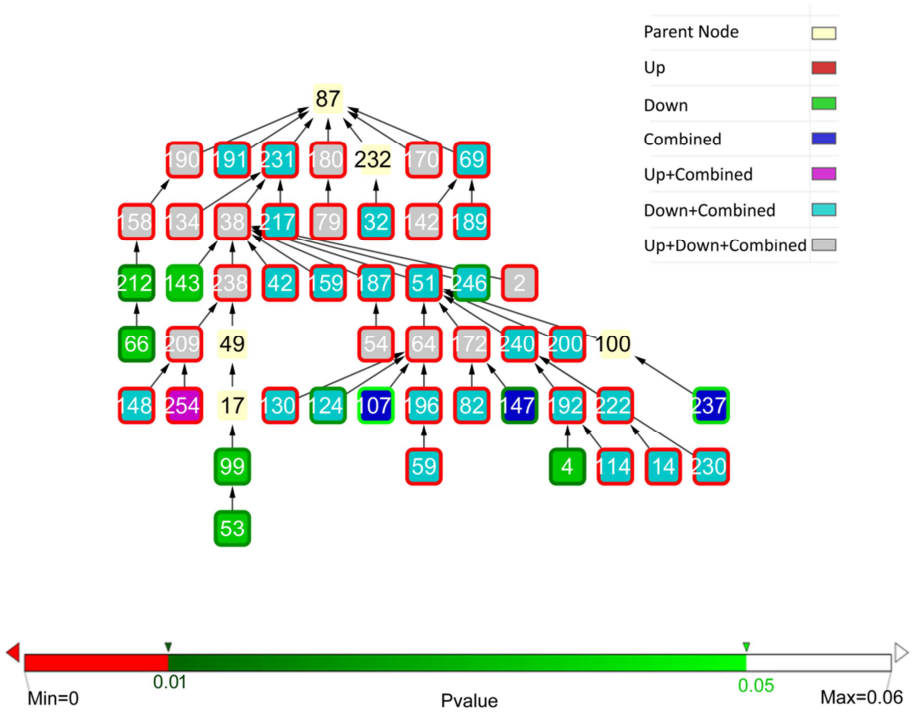


Figure 20. "Post-Translational Modification, Protein Folding, Drug Metabolism" gene regulatory network. The most related regulatory network (by Ingenuity Pathway Analysis®) for the Alarm_Pheromone genes that are conserved in all vertebrate species (in Inparanoid). Analysis is done using their human orthologs. 21 genes are involved in this network and they are all marked in red.

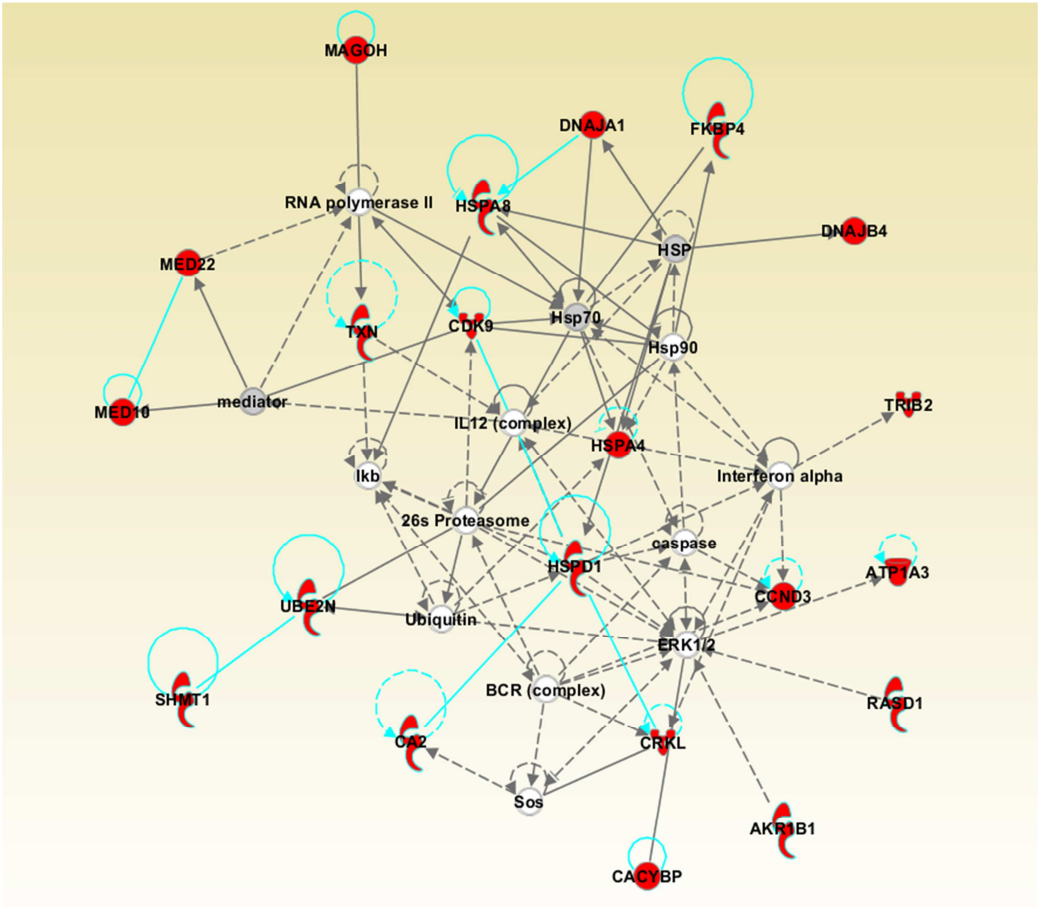


Figure 21. "Developmental Disorder, Hereditary Disorder, Neurological Disease" gene regulatory network. The 3rd most related regulatory network returned by Ingenuity Pathway Analysis for the Alarm_Pheromone genes that are conserved in all vertebrate species (in Inparanoid). Analysis is done using their human orthologs. 15 genes are involved in this network and they are all marked in red.

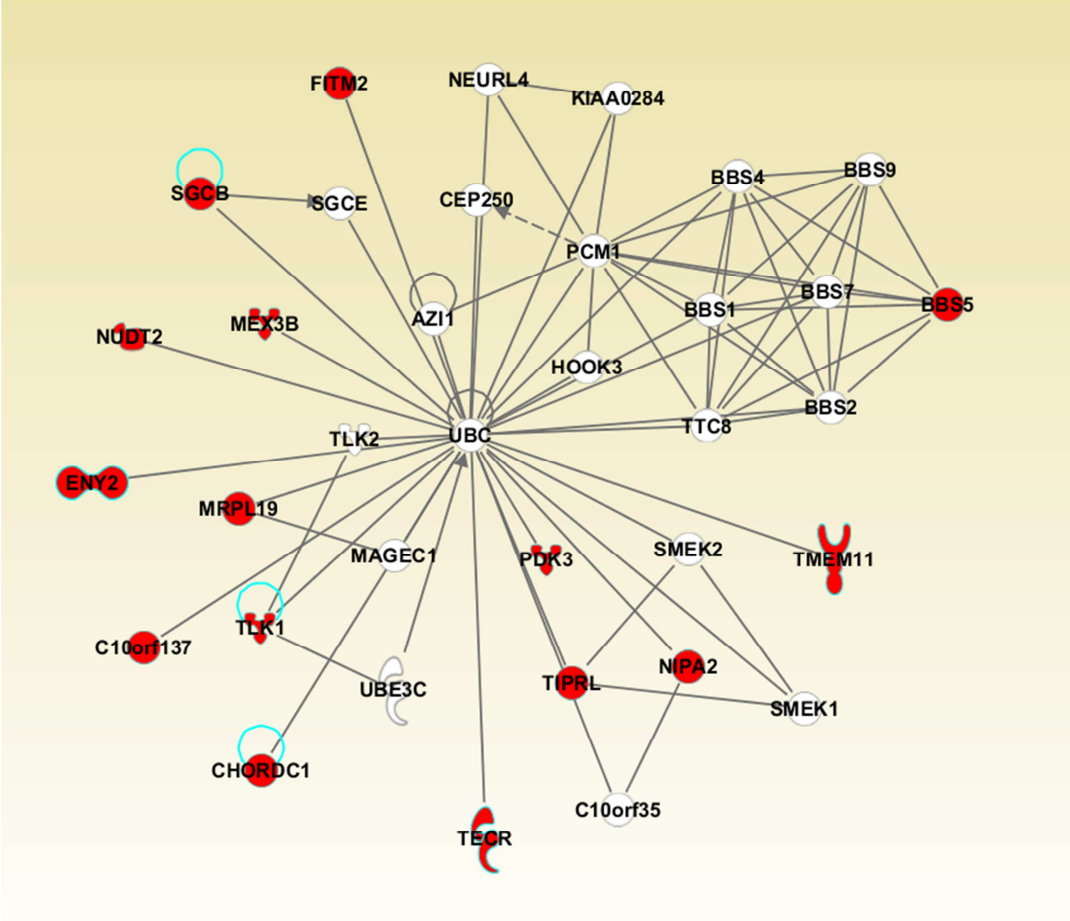


Figure 22. Co-regulation graph for the human orthologs of the 'Vertebrate-conserved' Alarm_Pheromone genes. The red nodes are human probe genes for which their bee orthologs were up-regulated in the Alarm_Pheromone dataset; purple nodes, down-regulated. The light blue and green nodes in the middle column were in the top 5 scoring genes for at least 2 and 3 probe genes, respectively. An arrow from a probe-node A to a node B means that B was in the top 5 scoring (D2z score) genes for probe A.

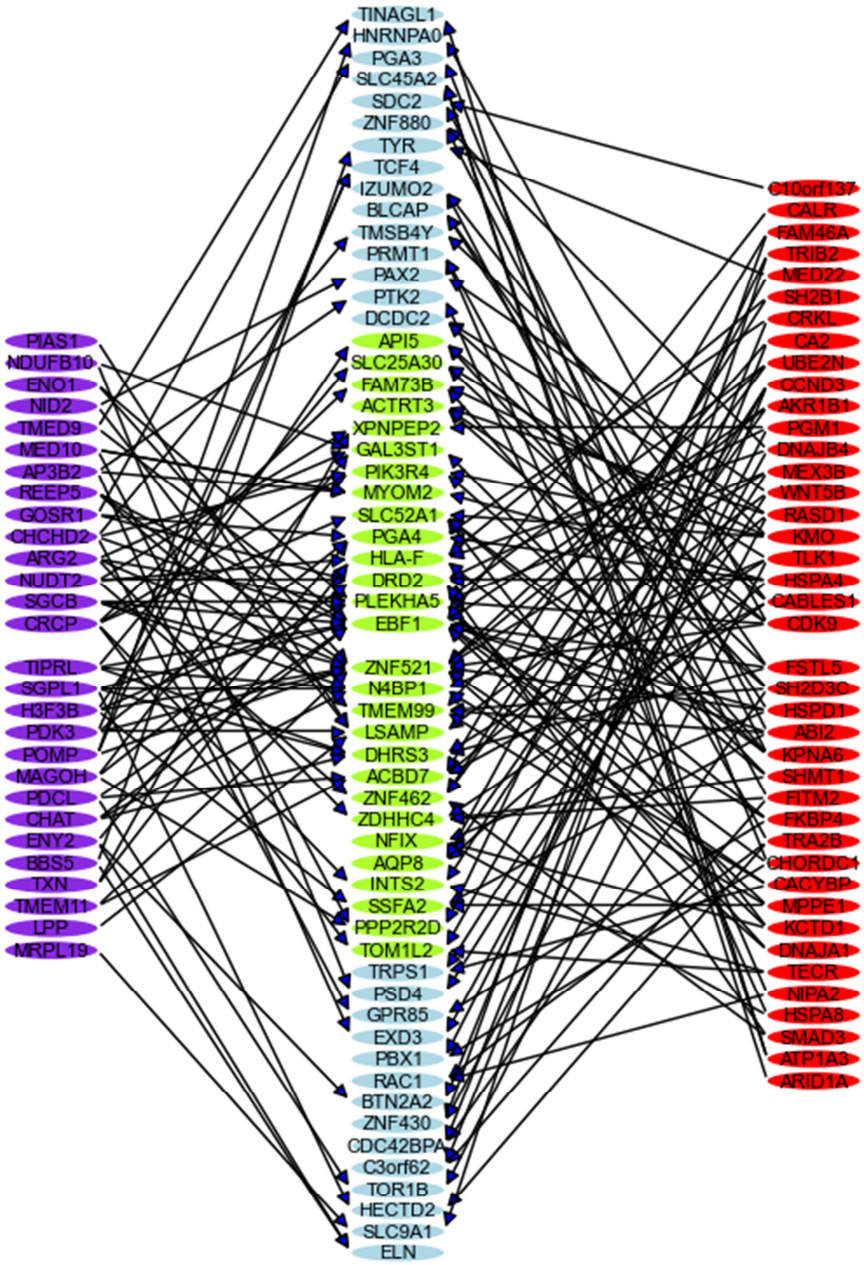


Figure 23. Co-regulation graph for the 'Mouse-human-conserved' Alarm_Pheromone Genes' Human Orthologs. 'Mouse-human-conserved' means that genes have orthologs in both mouse and human. The red (purple) nodes are human probe genes that their bee orthologs were up (down) regulated in the Alarm_Pheromone dataset. The light blue and green nodes in the middle column were in the top 5 scoring genes of at least 2 and 3 probe genes respectively. An arrow from a probe-node A to a node B means that B was in the top 5 scoring genes for probe A.

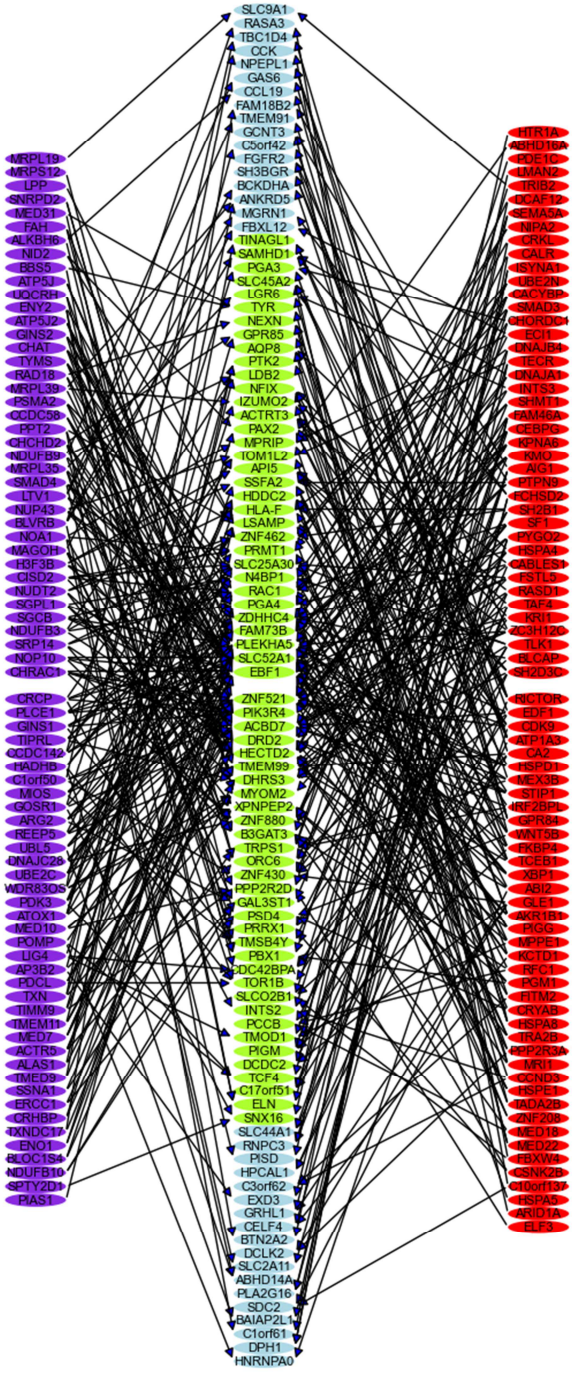


Figure 24. Co-regulation graph for the 'Mouse-human-conserved' up-regulated Alarm_Pheromone Genes' Human Orthologs. 'Mouse-human-conserved' means that genes have orthologs in both mouse and human. It is the same as Figure 23 except that it is based on only up-regulated Alarm_Pheromone genes. The light blue and green nodes in the middle column were in the top 5 scoring genes of at least 2 and 3 probe genes respectively.

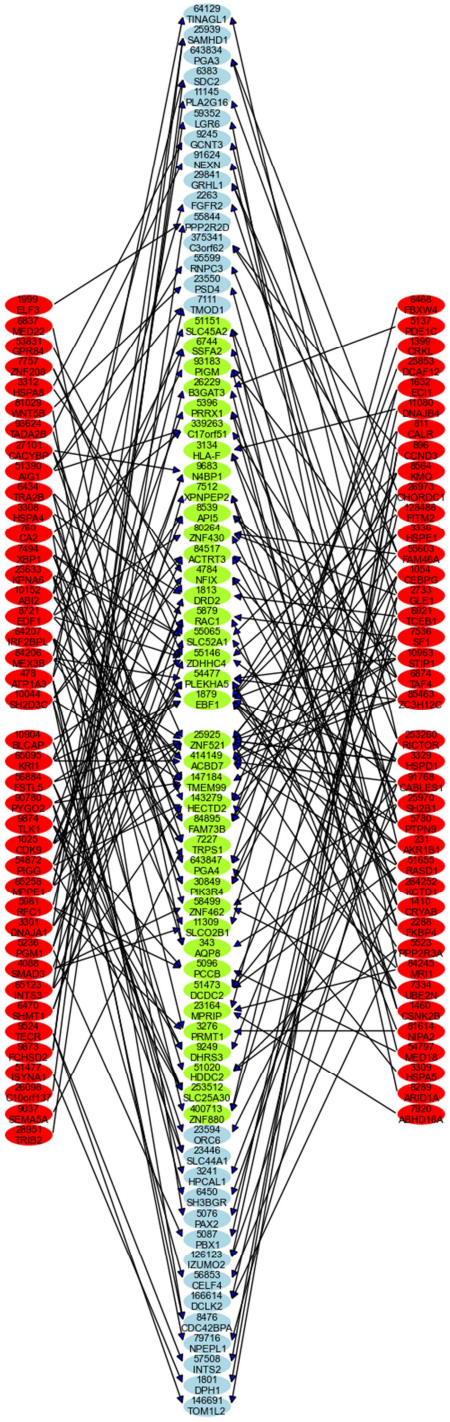


Figure 25. Co-regulation graph for the 'Mouse-human-conserved' down-regulated Alarm_Pheromone Genes' Human Orthologs. 'Mouse-human-conserved' means that genes have orthologs in both mouse and human. . It is the same as Figure 23 except that it is based on only down-regulated Alarm_Pheromone genes. The light blue and green nodes in the middle column were in the top 5 scoring genes of at least 2 and 3 probe genes respectively. An arrow from a probe-node A to a node B means that B was in the top 5 scoring genes for probe A.

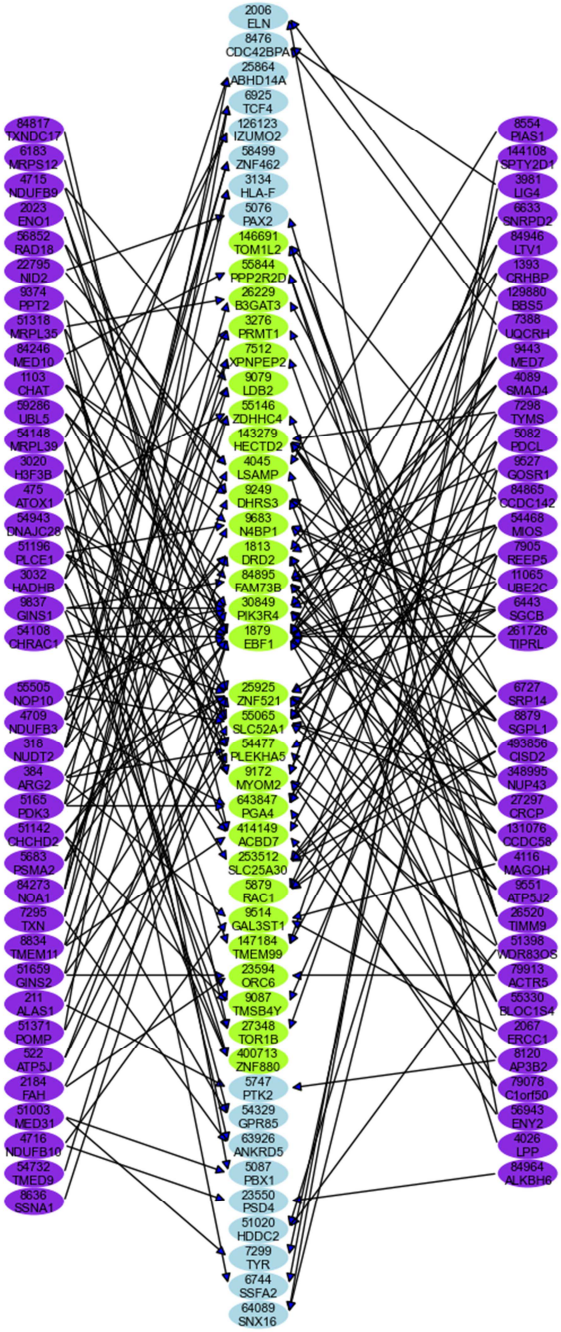


Figure 26. Example of hierarchical Optimal Multiple Alignment. Suppose we are given 8 motif sequences A-H, first we generated a guide tree as shown according to pairwise alignment. Second, considering the sequence lengths, we first did optimal multiple alignment of A,B,C together and produce a profile(A,B,C). Similarly, we aligned D,E together and produce a profile (D,E). And we aligned F,G,H together and produce a profile (F,G,H). Third, we can keep doing optimal multiple alignment of the profiles by aligning profile(A,B,C), profile(D,E) and profile(F,G,H) together.

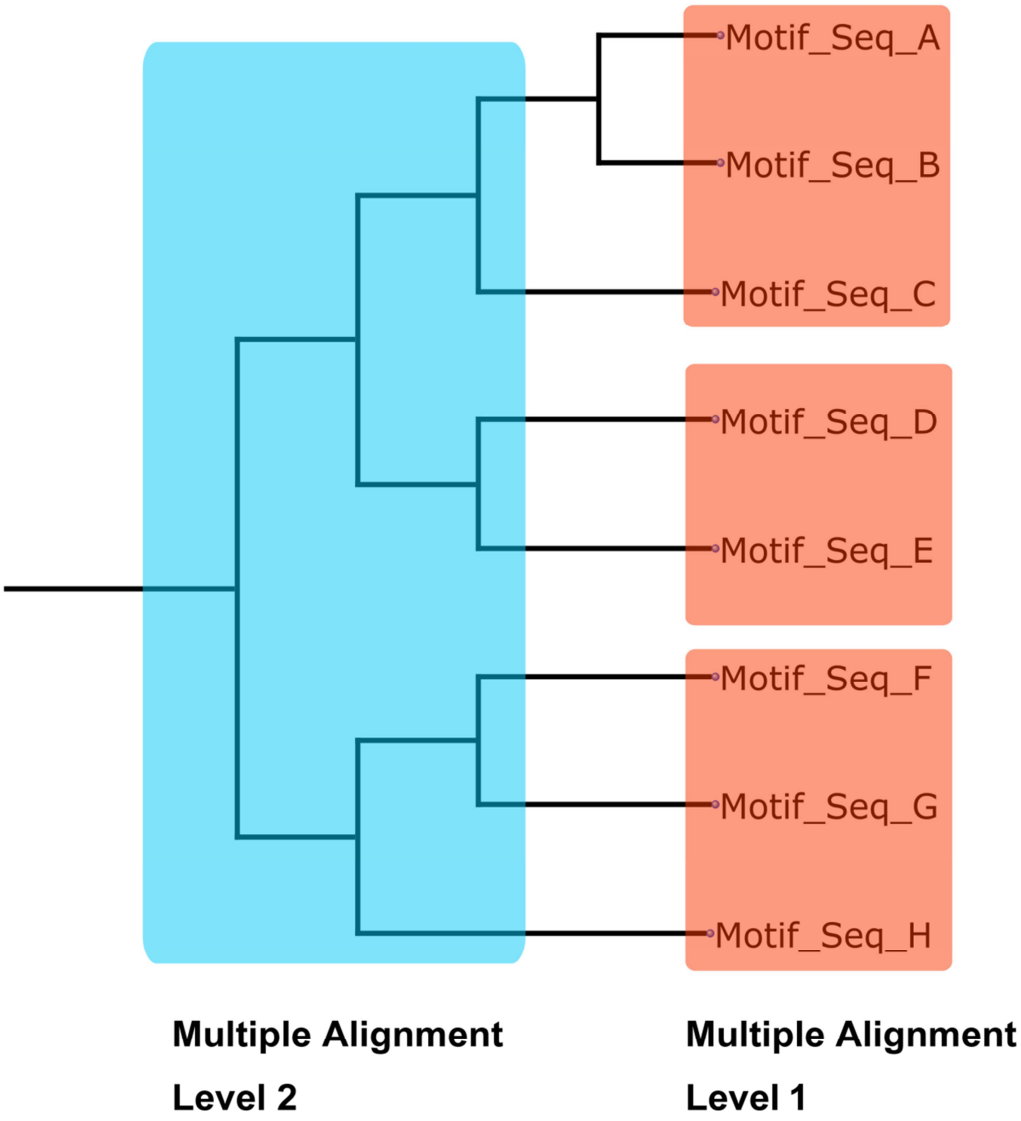
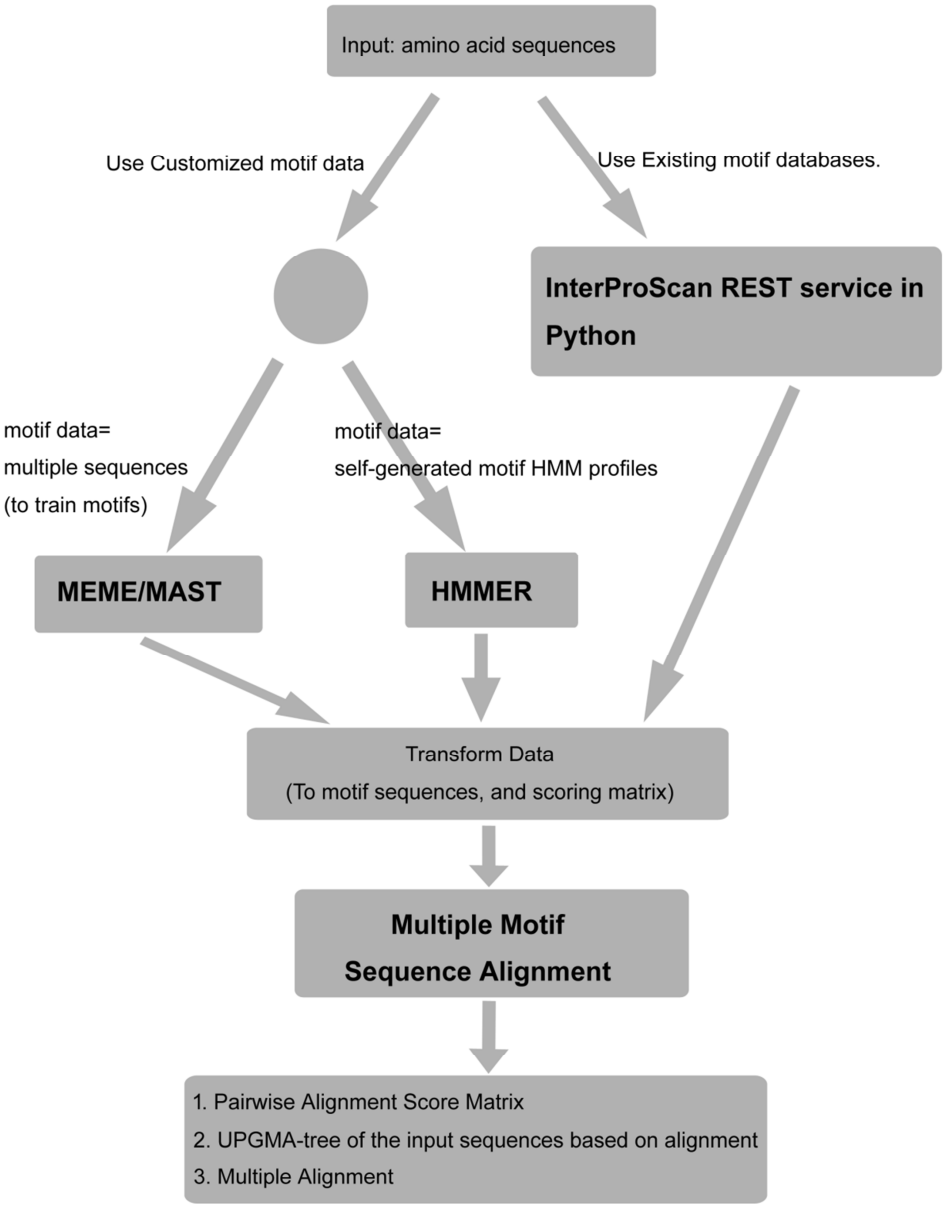


Figure 27. Schema of potential domain-based protein comparison pipeline. for each input sequence, a motif scanning tool is used: InteProScan(if we would like to use existing motif database), MEME/MAST(if we would like to discover novel motifs and we have a set of sequences for training), or HMMER(if have some self-generated motif Hidden-Markov-Model profiles that we would like to use.). Then motif/domain hit scores, and motif/domain compositions are transformed to scoring matrix and motif/domain sequences, which are input into Multiple Sequence Alignment tool. Finally, the multiple domain-sequence alignment, the pair-wise alignment score, and tree based on pair-wise alignment scores are generated.



APPENDIX A – Supplementary Tables

Table A1. Information of all the gene models in lizard, frog, chicken, zebra finch, opossum, mouse, dog and human. Table is divided into sections by species. Common columns are **'ID':** internal ID we assigned to each gene model;**'Evolution_Category':** defined according to the species that have ortholog of this gene by fingerprint alignment: 1) primate/dog/mouse/opossum/zebra_finch/chicken/frog/lizard (detected in human/dog/mouse/opossum/zebra_finch/frog/ lizard only, i.e species-specific); 2) eutherians (human and dog and/or mouse); 3) mammals (at least human and opossum); 4) amniotes (at least human and one bird or lizard); and 5) tetrapods (at least human and frog); **'List of Orthologous Species':** the actual species list that have ortholog of this gene by fingerprint alignment; **'Inferred_Gene_Type':** gene type inferred by its orthologs' type in the orthologous groups (see Table A2) defined by fingerprint alignment; **'Gene_Type':** the actual gene type based on the gene model; **'Annotation':** ENSEMBL or NCBI genes that overlaps with this gene model; **'Motif_Info':** the motif composition of the gene model, A means Krab_A motif, B means Krab_B motif, C means Krab_C motif, T means BTB/POZ motif, S means SCAN motif,Z means ZNF motif. For example, "ABZx15" means this gene model is composed of Krab_A, Krab_B, then followed by 15 ZNF motifs;**'Chromosome':** the chromosome this gene model resides on; **'Strand':** the DNA strand this gene model resides on; **'Start(1-based)':** start nucleotide number of the gene model,first nucleotide is numbered '1'; **'End(1-based)':** end nucleotide number of the gene model.For "**Human**" section we have two extra columns:'**Evolutionary pattern-curated':** manually curated evolutionary category;**'fingerprint conservation':**1=perfect conservation, 2=very high cons., 3= overall high conservation, 4=conserved but with several minor changes, 5=significant aa divergence, 6=significant indels, 7= aa changes and indels, 0= perfectly conserved but lost in mouse; For "**Mouse**" section we have three extra columns: **'ESTs of Mm9':** the EST data(from UCSC genome browser) overlapped with this gene model; **'Number of ESTs':** count of number ESTs overlapped;**'Potential Novel Gene':** Boolean variable indicating if the gene model is a potential new gene model: genes with no Ensembl model, but have EST overlap and at least 5 zinc fingers. There are 131 of them.

Table A2. The multiple 'fingerprint' alignments of orthologous groups of all gene models across 8 species. Orthologous groups are defined by cutting the hierarchical clustering tree as described in 'Methods' Section of Chapter 2. Each multiple alignment is started with a row starting with ">>>>", the number following the arrows is an empirical alignment score, the smaller the better, 0 is the score for perfect alignment.Second column of each alignment, are the IDs of the gene model started with species names. First column of each alignment are the annotation and gene model information available for the corresponding gene model. The fingerprint data of human and dog are from previous studies. Orthologous groups with single gene is kept as it is and no alignment is done.

Table A3. Result of Reciprocal Best Hits (RBHs) between human and mouse/chick/opossum/finch/dog/frog/lizard by both BLAST and fingerprint alignment. Reciprocal Best Hit (RBH) of fingerprint or BLAST is defined as in 'Methods' section of Chapter 2. As there could be multiple genes with same fingerprint alignment scores (or BLAST e-values), multiple fingerprint (or BLAST) RBHs could exist. Each cell in the table contains a list of RBHs separated by "|". Each RBH is listed as a name followed by a bracket containing a fingerprint alignment score (or a e-value and sequence coverage percentage). For Blast hits with e-value <= 1e-30, or Fingertip Hit with score <= 1.5(the score range is 0.0-2.0), they are colored as Good Hit "cyan". Otherwise, they are colored as Bad Hit "grey". For Human genes with no RBH in any species at all, they are also colored 'grey".

Table A4. Summary information about all the honey bee genes included in the 8 data sets studied. This table includes the following information of all the differentially expressed honey bee genes in this study: BeeBase ID, Refseq ID, Differential Expression Direction in each set (Only the signs of values are relevant. See Table 6 for how the experiments are set up. For example, for Alarm_Pheromone, when the number is positive that means that gene is up-regulated by alarm pheromone and vice versa.), ID on microarray, Entrez Gene ID, Gene Symbol, Gene Short Description, Whether Transcription Factor or not (based on Interpro Domain Information), Interpro Domain Composition, the ID's of their orthologs in different species (if present in InParanoid database, and "NA" if not) and total counts of orthologs(count 1 for each species)/homologs(count all homologs in each species).

Table A5. Result of the Gene Ontology enrichment analysis at p-value <=.05 level for human orthologs of 'Vertebrate-conserved' Alarm_Pheromone Genes. 'Vertebrate-conserved' here means the honey bee genes have orthologs in all vertebrate species of InParanoid (altogether 19, ranging from *T.nigroviridis* to *H.sapiens*). Their corresponding human orthologs are retrieved. Then Gene Ontology analysis is done on these human genes using DAVID by Entrez IDs (altogether 77).

Table A6. Result of the Gene Ontology enrichment analysis at p-value <=.05 level for human orthologs of all up-regulated 'Vertebrate-conserved' Alarm_Pheromone Genes. It is the same as Table A5 except that the analysis is done on up-regulated Alarm_Pheromone genes. Then Gene Ontology analysis is done on their human orthologs using DAVID by Entrez IDs (altogether 47).

Table A7. Result of the Gene Ontology enrichment analysis at p-value <=.05 level for human orthologs of all down-regulated 'Vertebrate-conserved' Alarm_Pheromone Genes. It is the same as Table A5 except that the analysis is done on down-regulated Alarm_Pheromone genes. Then Gene Ontology analysis is done on their human orthologs using DAVID by Entrez IDs (altogether 30).

Table A8. Result of the Gene Ontology enrichment analysis at p-value <=.05 level for human orthologs of all 'Mouse-human-conserved' Alarm_Pheromone Genes. 'Mouse-human-conserved' means genes have orthologs in human and mouse but not necessarily in other vertebrate species. Their corresponding human orthologs are retrieved. Then Gene Ontology analysis is done on these human genes using DAVID (altogether 169). All human genes are displayed by Entrez gene IDs.

Table A9. Result of the Gene Ontology enrichment analysis at p-value <=.05 level for human orthologs of all up-regulated 'Mouse-human-conserved' Alarm_Pheromone Genes. It is the same as Table A8 except that the analysis is done on up-regulated Alarm_Pheromone genes. Then Gene Ontology analysis is done on their human orthologs using DAVID (altogether 89). All human genes are displayed by Entrez gene IDs.

Table A10. Result of the Gene Ontology enrichment analysis at p-value <=.05 level for human orthologs of all down-regulated 'Mouse-human-conserved' Alarm_Pheromone Genes. It is the same as Table A8 except that the analysis is done on down-regulated Alarm_Pheromone genes. Then Gene Ontology analysis is done on their human orthologs using DAVID (altogether 80). All human genes are displayed by Entrez gene IDs.

Table A11. Ingenuity Pathway Analysis annotation of human orthologs of 'Vertebrate-conserved' Alarm_Pheromone Genes. 'Vertebrate-conserved' here means the honey bee genes have orthologs in all vertebrate species of InParanoid (altogether 19, ranging from *T.nigroviridis* to *H.sapiens*).Same as in Table A5.

Table A12. Tabulated results of the D2z analysis human orthologs of 'Vertebrate-conserved' Genes in each gene sets. This table shows the top 5 genes with most similar promoter region for each gene in all 8 gene sets analyzed. Every gene is the human gene in Entrez ID (and their synonyms are also listed if exist). Purple genes are down-regulated in original honeybee expression data. Red genes are up-regulated correspondingly. Green Genes are hubs hit more than 3 times. Light blue genes are hubs hit twice. 'Vertebrate-conserved' here means the honey bee genes have orthologs in all vertebrate species of InParanoid (altogether 19, ranging from *T.nigroviridis* to *H.sapiens*).

Table A13. Index tables for translation of GO term to index number for all GO terms appearing in GO trees. Top section is for GO analysis of human orthologs of Alarm_Pheromone set that has orthologs in 19 vertebrate species ('Vertebrate-conserved', Figure 12, 13). Bottom section is for Alarm_Pheromone set that has orthologs in mouse and human genomes but not necessarily in the other 17 vertebrates ('Mouse-human-conserved', Figure 14-19)."Set" column is defined the same as "Set Desc" in Table 9.

APPENDIX B – Supplementary Data Files

Datafile B1. "gene_models_6_species.zip". A zip file containing the gene model data built for lizard, frog, zebra finch, chicken, opossum and mouse. For each species, there are 5 files: ***.model.summary***: summary file of all the gene models generated according to 'Methods' section of Chapter 2, each row contains the following tab-delimited fields: **'ID'**: internal ID assigned to the gene model; 'chromosome': name of the chromosome the gene model resides on; **'strand'**: DNA strand the gene model resides on, + or -; **'gene type'**: motif_composition of the gene, for example TZx3 means that it contains a BTB/POZ motif followed by 3 ZNF motifs; **'start position(1-based)'**: start nucleotide of the gene model(with respect to + strand); **'end position'**: end nucleotide of the gene model(with respect to + strand); **'start codon'**: start codon used, or 'Startcodon' if no start codon can be found; **'stop codon'**: stop codon used; **'intron starts'**: comma-delimited list of start positions of introns(with respect to + strand); **'intron ends'**: comma-delimited list of end positions of introns(with respect to + strand); **'splicing sites'**: comma-delimited list of splicing site nucleotides for each introns, or 'CSEC' if intron is of zero length; **'motifs'**: comma-delimited list of motif IDs in order(with respect to + strand);**'fingerprints'**: comma-delimited list of fingerprints of each ZNF motif in order(with respect to + strand); **'ESTs'**: comma-delimited list of EST data overlapped with the gene model(ID is like 'EstGxxxx', the number after 'EstG' is the internal est group id used, which can be referred in Datafile B3), or 'NA' if no EST data overlaps with the gene model. ***.model.protein.fa***: fasta file of the protein sequences of each gene model in *.model.summary. ***_all_motifs.fa***: fasta file of peptide sequences of all motif hits returned by HMMER, stop codon in frame is annotated as "*". ***_all_motifs.gff***: gff file of all motif hits returned by HMMER, which can be uploaded to UCSC genome browser to be viewed. *_***.model.gff***: gff file of all gene models along with the motifs they contained, which can be uploaded to UCSC genome browser to be viewed(http://genome.ucsc.edu/FAQ/FAQformat.html#format3).

Datafile B2. "All.code". Tab-delimited file that contains the fingerprint sequences of all genes including human and dog (include even genes with single ZNFs). The uncertain fingerprint residue is annotated as 'x'.

Datafile B3. "ESTs_6_species.tar.gz". A compressed file that contains the EST files used for gene model building for frog,lizard,chick,finch,opossum and mouse. For each file *_*all_est.parsed.modified.grouped*, it is tab-delimited. Each row defines a EST group (EST fragments that overlaps with each other are grouped), and has the following 6 fields in order: id number(which is used in *.model.summary in Datafile B1), chromosome name, strand, start position(1-based),end position(1-based),list of EST ids(from UCSC genome browser) that belong to this EST group.